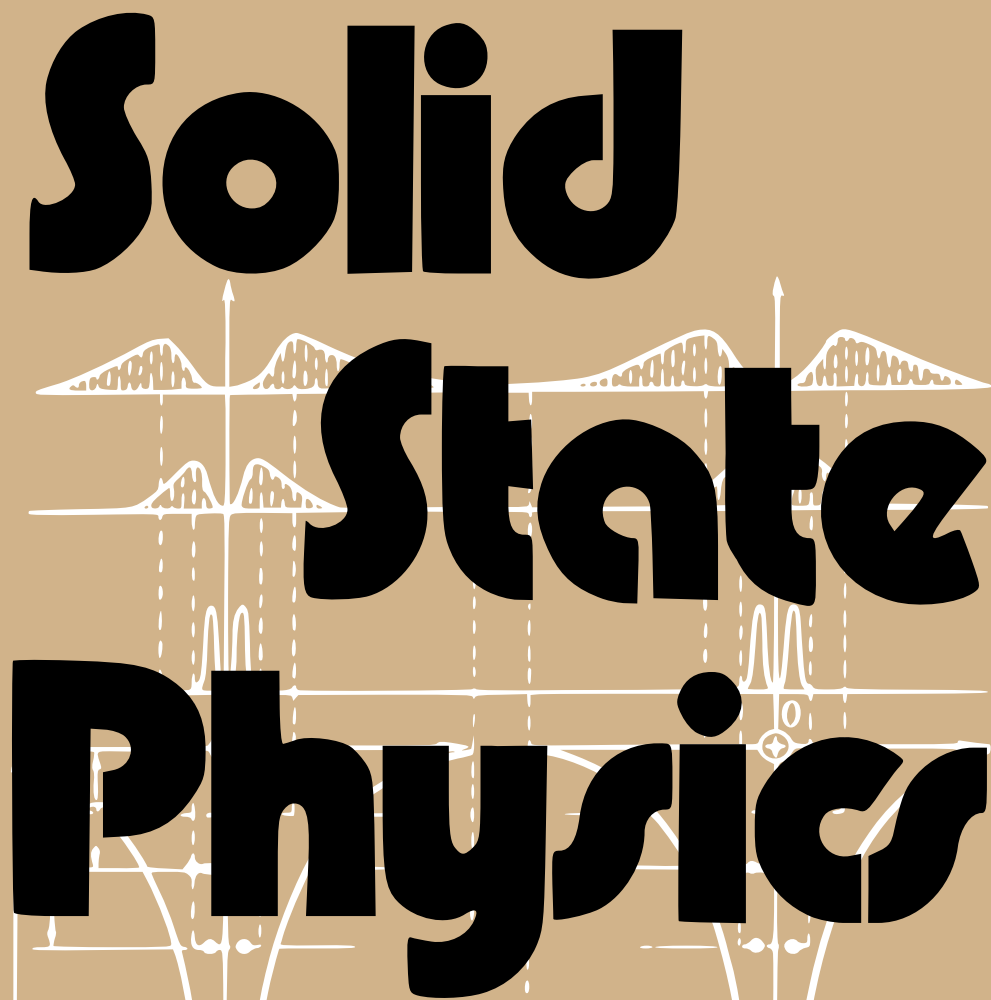
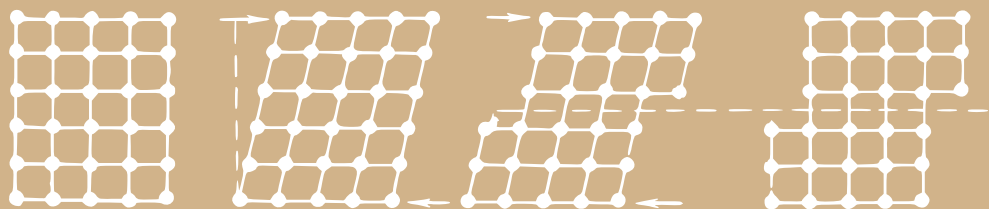


Solid State Physics

The background of the title section features several white line drawings on a tan background. At the top, there are four bell-shaped curves, each with a vertical dashed line extending downwards. Below these, there are two sets of oscillating waveforms, each with a vertical dashed line. In the center, there are two parabolic curves opening upwards, each with a vertical dashed line. To the right of the parables, there is a small circle with a cross inside, labeled with the number '0'. The overall theme is solid-state physics, specifically focusing on wave phenomena and lattice structures.

G. I. Epifanov



Mir Publishers Moscow

Solid State Physics

Solid State Physics

by **G. I. Epifanov**, D.Sc.

Moscow Institute of Electronic Engineering

Translated from the Russian by

Mark Samokhvalov, Cand. Sc.



MIR PUBLISHERS
MOSCOW

Revised from the 1977 Russian edition
First published 1979

English translation, Mir Publishers, 1979

PREFACE

Ten years have passed since the first Russian edition of this textbook was published. In this time solid state physics has developed rapidly as the scientific background of numerous front-line branches of technology, absorbing new discoveries and theories. This has been considered in preparing the new edition.

At the same time college curricula have been changed to improve the basic preparation of versatile engineers, especially in physics and mathematics. This too had to be reflected in this book.

Also, the years that have elapsed since the first edition have seen much comment, some critical, and many proposals from Soviet and foreign readers—from college teachers and students, teachers of vocational and secondary schools, engineers and scientists. The author is grateful for all the comment and proposals.

There was a need therefore to revise the book completely. As in the first edition, the presentation of material has followed the aim of elucidating the physical nature of the phenomena discussed. But, where possible, the qualitative relations are also presented, often though without rigorous mathematics.

The manuscript was reviewed in detail by Prof. L. L. Dashkevich, Dr. of Technical Sciences, and Prof. I. G. Nekrashevich, Honored Scientist of the Belorussian Republic. It was perused by Prof. L. A. Gribov, Dr. of Mathematical and Physical Sciences, Assistant Prof. V. B. Zernov, and Z. S. Sazonova. The author extends sincere thanks for their efforts and criticism, which he took into account when revising the manuscript.

The author is also indebted to Senior Lecturer F. Zh. Vilf, Cand. of Technical Sciences, and Assistant Prof. Yu. A. Moma, Cand. of Technical Sciences, for manuals used in this textbook on superconductivity, Gunn effect, and principles of operation of impulse and high-frequency diodes, and to Z. I. Epifanova for all her work in preparing the manuscript.

The author will be most grateful for comment and proposals that might improve this book. They should be sent to the publishers.

G. I. E.

Contents

Preface

vii

Chapter 1. Bonding. The Internal Structure of Solids

1

- § 1. The van der Waals forces 1
- § 2. The ionic bond 4
- § 3. The covalent bond 6
- § 4. The metallic bond 11
- § 5. The hydrogen bond 12
- § 6. Comparison between bonds of various kinds 14
- § 7. Forces of repulsion 15
- § 8. Crystal lattice 16
- § 9. Notation to describe sites, directions, and planes in a crystal 19
- § 10. Classification of solids based on the nature of bonds 22
- § 11. Polymorphism 27
- § 12. Imperfections and defects of the crystal lattice 30

Chapter 2. Mechanical Properties of Solids

35

- § 13. Elastic and plastic deformations. Hooke's law 35
- § 14. Principal laws governing plastic flow in crystals 40
- § 15. Mechanical twinning 44
- § 16. Theoretical and real shear strengths of crystals 45
- § 17. The dislocation concept. Principal types of dislocations 47
- § 18. Forces needed to move dislocations 51
- § 19. Sources of dislocations. Strengthening of crystals 53
- § 20. Brittle strength of solids 57
- § 21. Time dependence of the strength of solids 63
- § 22. Methods of increasing the strength of solids 67

Chapter 3. Elements of Physical Statistics

71

- § 23. Methods used to describe the state of a macroscopic system 71
- § 24. Degenerate and nondegenerate ensembles 75
- § 25. The number of states for microscopic particles 78
- § 26. Distribution function for a nondegenerate gas 81
- § 27. Distribution function for a degenerate fermion gas 83
- § 28. Distribution function for a degenerate boson gas 90
- § 29. Rules for statistical averaging 91

Chapter 4. Thermal Properties of Solids	95
§ 30. Normal modes of a lattice	95
§ 31. Normal modes spectrum of a lattice	98
§ 32. Phonons	99
§ 33. Heat capacity of solids	103
§ 34. Heat capacity of electron gas	108
§ 35. Thermal expansion of solids	110
§ 36. Heat conductivity of solids	114
Chapter 5. The Band Theory of Solids	121
§ 37. Electron energy levels of a free atom	121
§ 38. Collectivization of electrons in a crystal	124
§ 39. Energy spectrum of electrons in a crystal	126
§ 40. Dependence of electron energy on the wave vector	129
§ 41. Effective mass of the electron	134
§ 42. Occupation of bands by electrons	138
§ 43. Intrinsic semiconductors. The concept of a hole	140
§ 44. Impurity semiconductors	143
§ 45. Position of the Fermi level and free carrier concentration in semiconductors	146
§ 46. Nonequilibrium carriers	153
Chapter 6. Electrical Conductivity of Solids	157
§ 47. Equilibrium state of electron gas in a conductor in the absence of an electric field	157
§ 48. Electron drift in an electric field	157
§ 49. Relaxation time and mean free path	159
§ 50. Specific conductance of a conductor	161
§ 51. Electrical conductivity of nondegenerate and degenerate gases	162
§ 52. Wiedemann-Franz-Lorenz law	164
§ 53. Temperature dependence of carrier mobility	165
§ 54. Electrical conductivity of pure metals	171
§ 55. Electrical conductivity of metal alloys	172
§ 56. Intrinsic conductivity of semiconductors	176
§ 57. Impurity (extrinsic) conductivity of semiconductors	177
§ 58. Deviation from Ohm's law. The effect of a strong field	180
§ 59. The Gunn effect	182
§ 60. Photoconductivity of semiconductors	183
§ 61. Luminescence	190
§ 62. Fundamentals of superconductivity	194
Chapter 7. Magnetic Properties of Solids	211
§ 63. Magnetic field in magnetic materials	211
§ 64. Magnetic properties of solids	212
§ 65. Magnetic properties of atoms	219
§ 66. Origin of diamagnetism	224
§ 67. Origin of paramagnetism	227
§ 68. Origin of ferromagnetism	233
§ 69. Antiferromagnetism	241

§ 70. Ferrimagnetism. Ferrites	242
§ 71. Magnetic resonance	243
§ 72. Fundamentals of quantum electronics	245
Chapter 8. Contact Phenomena	251
§ 73. Work function	251
§ 74. Contact of two metals	254
§ 75. The metal-semiconductor contact	257
§ 76. Contact between two semiconductors of different types of conductivity	264
§ 77. Physical principles of semiconductor p-n junction devices	274
§ 78. Fundamentals of integrated circuit electronics (microelectronics)	284
Chapter 9. Thermoelectric and Galvanomagnetic Phenomena	289
§ 79. The Seebeck effect	289
§ 80. The Peltier effect	294
§ 81. The Thomson effect	297
§ 82. Galvanomagnetic phenomena	297
§ 83. Practical applications of thermoelectric and galvanomagnetic phenomena	302
APPENDICES	305
A.1. Derivation of the Maxwell-Boltzmann distribution function	305
A.2. Derivation of the Fermi-Dirac distribution function	306
A.3. Derivation of the Bose-Einstein distribution function	308
A.4. Tables	309
Glossary of Symbols and Notations	311
Bibliography	317

Chapter 1

Bonding. The Internal Structure of Solids

Matter can exist in the solid state only because there are forces of interaction acting between the structural particles when the latter are brought sufficiently close together. For the solid to have a stable structure the forces of interaction between the particles should be of two types: of attraction, to prevent the particles from moving away from each other, and of repulsion, to prevent the particles from merging.

Let us discuss briefly the nature of these forces.

§ 1. The van der Waals forces

The most general type of bond existing between any two atoms, or molecules, is due to van der Waals forces. Those forces were first introduced to explain the equation of state of real gases, the *van der Waals equation*:

$$\left(p + \frac{a}{V^2}\right)(V - b) = RT \quad (1.1)$$

where the correction terms a/V^2 and b account, respectively, for the effect of the forces of attraction and repulsion acting between the molecules of a real gas. These forces manifest themselves in an almost ideal form in the interaction between the molecules with saturated chemical bonds (O_2 , H_2 , N_2 , CH_4 , etc.), as well as between the atoms of inert gases, making possible their existence in the liquid and solid states.

As a general case, the van der Waals bond is made up of the dispersion, orientational and induction interactions. Let's consider them separately.

Dispersion interaction. Consider the simplest example of two interacting helium atoms shown in Figure 1.1. The electron density distribution in a helium atom is spherically symmetrical and for this reason its electric moment is zero.

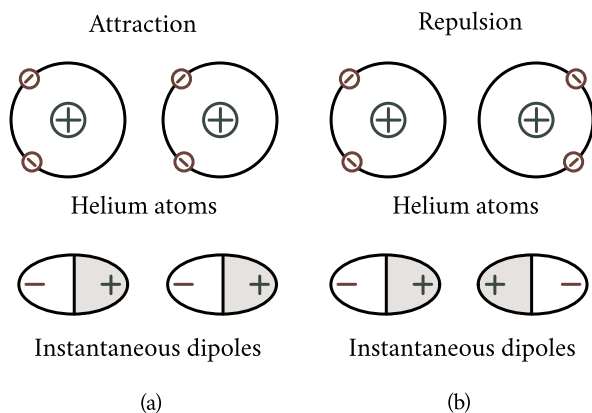


Figure 1.1: Dispersion interaction. The interaction between helium atoms is due to the correlation in the motion of electrons resulting in the appearance of instantaneous dipoles: (a) — correlation resulting in attraction; (b) — correlation resulting in repulsion.

But this means only that the average electric moment of the atom is zero. At every moment of time the electrons occupy particular points in space, thereby creating instantaneous rapidly changing electric dipoles. When two helium atoms are brought together, the motion of their electrons becomes correlated and this leads to the forces of interaction. The forces can be of two types. If the motion of the electrons is correlated, as shown in Figure 1.1(a), the instantaneous dipoles attract each other; if the correlation is as shown in Figure 1.1(b), the resulting interaction is repulsion. Since the realization of the arrangement of Figure 1.1(a) leads to a reduction in the energy of the system, this arrangement is more probable and is realized more frequently. This is in effect the cause of the constantly existing force of attraction between helium atoms.

The bonds discussed above that owe their existence to a correlation in motion of the electrons in adjacent atoms are termed *dispersion forces*. They were first calculated by F. London in 1930. The calculation was based on the following model: the instantaneous electric dipole of one atom causes the other atom to be polarized and it becomes an induced dipole leading to the realization of the arrangement in Figure 1.1(a), which corresponds to attraction. The calculation had as its final result the following expression for the energy of the dispersion interaction of two particles:

$$U_d = -\frac{3}{4} \frac{\alpha^2 E_{\text{exc}}}{r^6} \quad (1.2)$$

where α is the polarizability of the particles¹, E_{exc} their energy of excitation, and r

¹Let us recall the physical meaning of α . The charges in the molecule are displaced under the

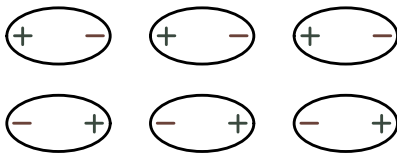


Figure 1.2: Orientational interaction of polar molecules.

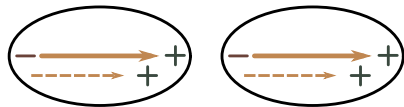


Figure 1.3: Induction interaction of molecules (dashed lines show the induced dipoles).

the distance between them.

Orientational interaction. Should the molecules possess a constant dipole moment M , that is, should they be polar molecules, an electrostatic interaction would be established between them tending to arrange the molecules in a strict order (Figure 1.2), since that order corresponds to the minimum energy of the system. The correct orientation of the molecules is disturbed by thermal motion. Therefore, the energy of the system due to the mutual orientation of the molecules is strongly dependent on temperature. At low temperatures, when the orientation of the molecules is perfect, the interaction energy is determined from the expression

$$U_{\text{or}} = -\frac{M^2}{2\pi\epsilon_0 r^3} \quad (1.3)$$

where r is the distance between the molecules, and ϵ_0 the permittivity of free space.

In the high temperature range the energy of interaction of polar molecules, as had been demonstrated by W. H. Keesom, is of the following form:

$$U_{\text{or}} = -\frac{M^4}{24\pi^2\epsilon_0^2 k_B T r^6}. \quad (1.4)$$

where k_B is the Boltzmann constant and T the temperature.

The type of interaction discussed above is termed *orientational interaction*.

Induction interaction. Lastly, in case of polar molecules of high polarizability an induced moment due to the action of the field of constant dipoles may be established (Figure 1.3; the induced dipoles are shown by dashed lines). The energy of mutual attraction due to the interaction of the rigid dipole of the first molecule and the induced dipole of the second molecule, as has been shown by Debye, is independent of temperature and is given by the expression

$$U_{\text{in}} = -\frac{\alpha M^2}{8\pi\epsilon_0^2} \frac{1}{r^6} \quad (1.5)$$

where, as before, M is the constant dipole moment of the molecule, and α its po-

action of an external field of intensity \mathcal{E} . This leads to a dipole moment M proportional to \mathcal{E} : $M = \alpha\mathcal{E}$, the proportionality factor α being the *polarizability* of the molecule.

larizability.

Such interaction is termed *induction*, or *deformation*, interaction.

In general, when two molecules are brought close together all three types of interaction may be established, the interaction energy being the sum of the energies of the dispersion (U_d), orientational (U_{or}), and induction (U_{in}) types of interaction:

$$U = U_d + U_{or} + U_{in}.$$

Table 1.1 shows the relative magnitude (in percent) of each of those components of the total bonding energy for water, ammonia, hydrogen chloride and carbon monoxide. The data presented in Table 1.1 show the induction interaction for all the substances to be weak. Three quarters or a half of the bond energy in substances made up of polar molecules is due to the energy of orientational interaction; while in materials made up of nonpolar molecules almost all of the bond energy is due to the dispersion interaction.

Table 1.2 shows the values of the bond energy for some molecular crystals held together by van der Waals forces.

§ 2. The ionic bond

Atoms that occupy places in the Mendeleev periodic table next to inert gases tend to assume the electronic configuration of these gases either by giving away or accepting electrons. The valence electron of alkali metals, which immediately follow the inert gases, moves outside the closed shell and is only weakly connected with the nucleus. The halides, which immediately precede the inert gases, lack one electron to complete a stable shell characteristic of an inert gas. Therefore, they exhibit high affinity to an excess electron.

Such atoms, that is, typical metals and halides, are bonded in the following way.

Table 1.1

Substance	Type of interaction		
	Dispersion	Induction	Orientalional
Water	19	4	77
Ammonia	50	5	45
Hydrogen chloride	81	4	15
Carbon monoxide	100		

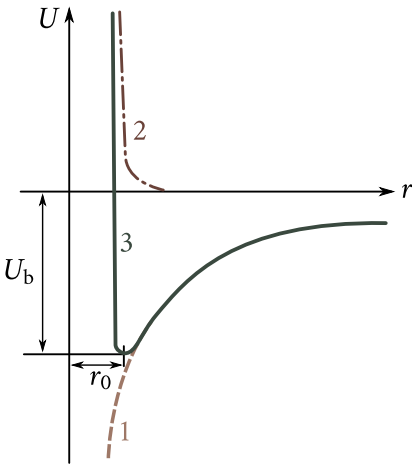


Figure 1.4: Dependence of energy of interacting ions on the distance between them: 1 — energy of attraction, 2 — energy of repulsion, 3 — total energy of interaction.

First a recharging of the atoms takes place. The electron from the metal moves over to the halide. This turns the metal atom into a positively charged ion and the haloid atom into a negatively charged one. These ions interact according to the Coulomb law as two opposite charges. Such a bond became known as an *ionic bond*.

The energy of attraction of two ions separated by the distance r is

$$U_{\text{att}} = -\frac{q^2}{4\pi\epsilon_0 r} \tag{1.6}$$

where q is the charge of the ions.

The curve 1 in Figure 1.4 shows the dependence of U_{att} on r . As r decreases the

Table 1.2

Substance	$U_b, [10^3 \text{ J mol}^{-1}]$
Neon (Ne)	1.90
Argon (Ar)	8.40
Nitrogen (N ₂)	6.60
Carbon monoxide (CO)	8.40
Oxygen (O ₂)	8.20
Methane (CH ₄)	10.8

absolute value of the energy increases monotonically, tending to infinity as $r \rightarrow 0$. The force of attraction tends to bring the ions together as close as possible. This, however, is prevented by the forces of repulsion, which begin to make themselves felt at small distances and rise very rapidly with the decrease in distance. The repulsion energy U_{rep} is shown in Figure 1.4 by the curve 2. Max Born and other scientists expressed the repulsion energy in the form

$$U_{\text{rep}} = \frac{B}{r^n} \quad (1.7)$$

where B and n are constants.

The resulting interaction energy of two ions is

$$U = U_{\text{rep}} + U_{\text{att}} = \frac{B}{r^n} - \frac{q^2}{4\pi\epsilon_0 r}. \quad (1.8)$$

This energy is shown in Figure 1.4 by the curve 3 which has a minimum at $r = r_0$; the depth of this minimum determines the bond energy U_b , and r_0 determines the distance between the ions in the molecule. Making use of the fact that in equilibrium (at $r = r_0$) the force of attraction, $F_{\text{att}} = -(dU_{\text{att}}/dr)_{r=r_0}$, equals the force of repulsion, $F_{\text{rep}} = -(dU_{\text{rep}}/dr)_{r=r_0}$, we can easily express (1.8) as

$$U_b = -\frac{q^2}{4\pi\epsilon_0 r_0} \left(1 - \frac{1}{n}\right). \quad (1.9)$$

The energy of the lattice made up of N such molecules is

$$U_{\text{lattice}} = -NA \frac{q^2}{4\pi\epsilon_0 r_0} \left(1 - \frac{1}{n}\right) \quad (1.10)$$

where A is the *Madelung constant*, which takes account of the energy of interaction of the given molecule with all its neighbouring molecules in the crystal.

Table 1.3 shows by way of an example the experimental values of the bond energy of some ionic crystals and its values calculated with the aid of (1.10). The discrepancies do not exceed 1-2 percent, which is proof of good agreement between theory and experiment.

§ 3. The covalent bond

The ionic and van der Waals bonds are unable to account for the existence of such compounds as H_2 , O_2 , N_2 , etc., as well as for bonds in atomic crystals of the diamond type. Evidently, atoms of one kind cannot form oppositely charged ions by changing the distribution of valence electrons, as was the case in the metal-halide interaction. On the other hand, the bond in the H_2 , O_2 and N_2 molecules is much stronger than that which could be attributed to the van der Waals forces. For such compounds the role of the van der Waals forces is that of a small correction to the

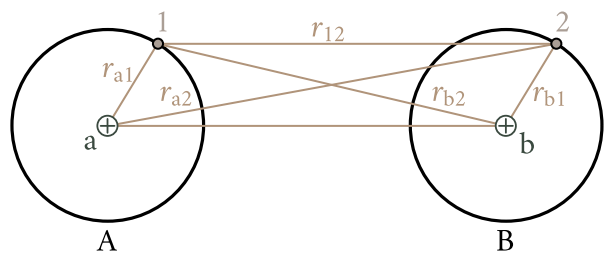


Figure 1.5: Calculating covalent bond between hydrogen atoms: A, B — hydrogen atoms; a, b — their nuclei; 1 — electron of atom A; 2 — electron of atom B; r_{a1} , r_{b2} — distances of electrons from their nuclei; r_{12} — distance between electrons; r_{a2} — distance of electron 2 from nucleus a; r_{b1} — distance of electron 1 from nucleus b; r — distance between nuclei.

bond mainly responsible for the strength of the compounds. This bond became known as the *covalent bond*.

Let us consider the nature of this type of bond using the hydrogen molecule as an example.

Suppose that two hydrogen atoms are at a rather great distance r from one another. The atom A consists of the nucleus a and the electron 1 and the atom B consists of the nucleus b and the electron 2 (Figure 1.5). Since the density of the electron cloud which describes the electron state in an atom falls off very rapidly as the distance from the nucleus increases, the probabilities to discover electron 1 near nucleus b and electron 2 near nucleus a are very small. Calculation shows that for $r \approx 50 \text{ \AA}$ each electron can visit the other nucleus on the average only once in 10^{12} years. Because of that atoms A and B may be regarded as isolated atoms and the energy of the system made up of such atoms may be taken to be equal to $2E_0$, where E_0 is the energy of the isolated atom in the ground state.

When the atoms are brought closer together, the probability of the electrons

Table 1.3

Crystal	$U_b, [10^3 \text{ J mol}^{-1}]$	
	Experiment	Theory
Sodium chloride (NaCl)	752	754
Potassium iodine (KI)	650	630
Rubidium bromide (RbBr)	635	645
Caesium iodine (CsI)	595	585

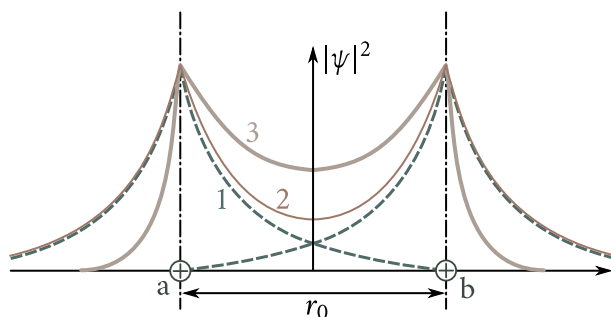


Figure 1.6: Electron density distribution in a system of two hydrogen atoms: 1 — electron density distribution in isolated hydrogen atoms, 2 — electron density resulting from simple overlapping of electron clouds of isolated atoms brought together to within a distance of r , 3 — actual electron density distribution in a hydrogen molecule.

going over to nuclei other than their own increases. For $r \approx 2 \text{ \AA}$ the electron clouds begin to overlap noticeably and the transition frequency rises up to 10^{14} s^{-1} . If the atoms are brought still closer together, the frequency of the electron exchange rises so that it becomes meaningless to speak of electron 1 as belonging to the atom A and of electron 2 as belonging to atom B. This corresponds to a new state that is not characteristic for a system made up of two isolated atoms. A remarkable property of this new state is that the electrons in it belong simultaneously to both nuclei, in other words, are *collectivized*.

The collectivization of the electrons is accompanied by a change in the electron density distribution $|\psi|^2$ and in the energy of the system as compared to the total energy $2E_0$ of the isolated atoms. The lines 1 in Figure 1.6 show the electron cloud density of the isolated atoms, the line 2 shows the total density obtained by simple superposition of the electron clouds of isolated atoms, and the line 3 the actual density distribution along the axis joining the nuclei a and b brought about by the collectivization of the electrons. The figure shows that the collectivization of the electrons results in the electron clouds being drawn into the space between the two nuclei: at a small distance from the nucleus outside this space the density of the clouds falls off, as compared with the density in isolated atoms, at the same time rising in the space between the nuclei above the sum of the densities of isolated atoms. The appearance of a state with an electron cloud of increased density that fills the space between the nuclei always results in a decrease in the system's energy and in the appearance of forces of attraction between the atoms. Speaking figuratively, we may say that the electron cloud formed in the space between the nuclei by a collectivized pair of electrons draws the nuclei together, striving to bring them as close together as possible.

Such is the qualitative picture of the origin of the covalent bond. Quantitative calculations of the hydrogen molecule were first performed by W. H. Heitler and F. London in 1927. Those calculations have shown that a system consisting of two closely spaced atoms of hydrogen can have two energy values depending on the direction of the electron spins in the atoms:

$$U_s = 2E_0 + \frac{(K + A)}{(1 + S^2)} \quad (1.11)$$

when the spins are antiparallel, and

$$U_a = 2E_0 + \frac{(K + A)}{(1 - S^2)} \quad (1.12)$$

when the spins are parallel. Here $2E_0$ is the total energy of the two isolated atoms, K is the *energy of the electrostatic interaction* of the electrons with the nuclei, of the electrons with one another, and of the nuclei. Another name for it is *Coulomb energy*. Its sign is negative. By A we denote the *energy of exchange interaction* due to the atoms exchanging electrons. This is the additional energy that appears as the result of the change in the electron density distribution in the process of the formation of the molecule. Its sign is negative and its absolute value is much larger than K ($|A| \gg |K|$). S is the *overlap integral* whose value lies within the limits $0 \leq S \leq 1$.

The state with the energy U_s is termed *symmetric* and with U_a *antisymmetric*. Since both K and A are negative and $S \leq 1$, the energy of the system in the symmetric state is less than the energy of two isolated atoms:

$$U_s < 2E_0. \quad (1.13)$$

This corresponds to the appearance of forces of attraction. Since the absolute value of the exchange energy A is considerably greater than that of the Coulomb energy K the decrease in the system's energy is mainly due to A . For this reason the force of attraction that appears between the atoms is termed the *exchange force*.

For the same reason, that is, because $|A| \gg |K|$, the formation of the antisymmetric state leads to an increase in the system's energy. This corresponds to the appearance of repulsive forces.

Figure 1.7 shows the dependence of U_s and U_a on r/a , where r is the distance between the atoms, and $a = 0.529 \text{ \AA}$ is the radius of the first Bohr orbit (the *Bohr radius*). The zeroth energy level has been fixed at $2E_0$. Figure 1.7 shows that in the antisymmetric state the system's energy rises steadily as the atoms are brought closer together (curve 7), this corresponding to the mutual repulsion of the atoms. Therefore a hydrogen molecule cannot be formed in such a state. In the symmetric state (curve 2) the system's energy at first falls as the distance r between the atoms decreases, attaining its minimum value at $r = r_0$. As the distance r decreases

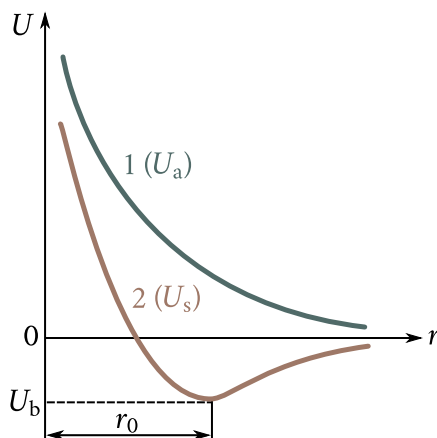


Figure 1.7: Interaction energy of two hydrogen atoms: 1 — antisymmetric state, 2 — symmetric state.

still further, the energy begins to rise because of the appearance of strong repulsive forces. The existence of a minimum on the potential energy curve makes the existence of a stable system of two hydrogen atoms, that is, a hydrogen molecule, possible. To destroy this system work must be performed equal to the depth of the potential well, U_b .

Calculation provides the following values of U_b and r_0 : $U_b = 4.37 \text{ eV}$, $r_0 = 0.735 \text{ \AA}$; the experimental values are $U_b = 4.38 \text{ eV}$, $r_0 = 0.753 \text{ \AA}$. The agreement between theory and experiment is quite good.

Table 1.4 shows the values of the bond energy for some covalent compounds—the molecules of H_2 , N_2 , O_2 , CO —and for diamond, silicon and germanium crystals in which the bonding is due to covalent forces.

Characteristic properties of the covalent bond, which distinguish it from the bonds of other types, are its *saturability* and *directionality*.

Saturability means that each atom can form covalent bonds only with a limited number of its neighbours. This means that each hydrogen atom can form covalent bonds only with one of its neighbours. The electron pair constituting such a bond has antiparallel spins and occupies one quantum cell. A third atom in this case instead of being attracted will be repelled.

The valence bond is formed in the direction of the greatest density of the electron cloud corresponding to the valence electrons. In this case there is maximum overlapping of the electron clouds of the bonding electrons, which implies that the valence bond is directional.

§ 4. The metallic bond

There is a special group of substances, called metals, that occupy places at the beginning of every period of the Mendeleev table. The formation of the metallic bond cannot be explained by the presence of the ionic or the covalent bond. Indeed, the ionic bond appears only between atoms having different affinities to the additional electron, for instance, between the atoms of a metal and a halide. Evidently, such bond between kindred atoms of a metal having identical affinity to the electron is impossible. On the other hand, metallic atoms do not have enough valence electrons to form valence bonds with their nearest neighbours. For instance, the copper atom has one valence electron and can form a valence bond only with a single atom. But in the copper lattice every atom is surrounded by twelve neighbours with which it must be connected by lines of force. This points to the fact that in metals there is a special type of bonding known as the *metallic bond*. Let us consider the nature of this bond.

In metallic atoms the external valence electrons are rather weakly coupled to the nucleus. In the liquid and solid states the atoms come so close together that the valence electrons are able to leave their respective atoms and wander throughout the lattice. This leads to an extremely homogeneous distribution of the negative charge in the crystal lattice. This conclusion is supported by direct experiments. Figure 1.8 shows an experimental curve of the electron density distribution between the sites of the aluminium lattice obtained by means of X-ray photography. Most part of the distance between the sites the electron concentration remains constant. Only quite close to the sites it rises sharply because of the presence of

Table 1.4

Gas	$U_b, [10^5 \text{ J mol}^{-1}]$
Carbon monoxide (CO)	10.8
Nitrogen (N ₂)	9.50
Oxygen (O ₂)	5.00
Hydrogen (H ₂)	4.40
Diamond (C)	6.80
Silicon (Si)	4.40
Germanium (Ge)	3.50

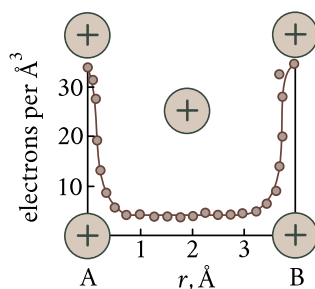


Figure 1.8: Electron density distribution in the aluminium lattice obtained by X-ray photography.

internal shells of the aluminium atoms.

In the lattice of a metal the bond is due to the interaction of the positive ions with the electron gas. The electrons moving between the ions compensate the forces of repulsion existing between the positively charged ions and draw them closer together. As the distance between the ions becomes smaller the density of the electron gas rises and this leads to an increase in the force drawing the ions together. On the other hand, in this case the repulsive forces acting between the ions tend to move them away from each other. When the distance between the ions becomes such that the forces of attraction are compensated by the forces of repulsion, a stable lattice is formed.

It appears that the metallic bond is somewhat similar to the valence bond, since they are both based on the collectivization of external valence electrons. However, in case of the valence bond only atoms that form pairs of nearest neighbours take part in the collectivization of electrons, and the respective electrons always remain between the atoms. In case of the metallic bond all atoms of the crystal take part in the collectivization of electrons, and the collectivized electrons are no longer localized near their respective atoms but move freely inside the lattice.

§ 5. The hydrogen bond

The *hydrogen bond* is formed between an atom of hydrogen and an extremely electronegative atom, for instance, an atom of oxygen, fluorine, nitrogen, chlorine. Such an atom attracts the bonding electrons and becomes negatively charged; the hydrogen atom after losing the bonding electron assumes a positive charge. The hydrogen bond is a result of electrostatic attraction of those charges.

As a typical example we may cite the hydrogen bond between molecules of water (Figure 1.9). The O–H bond between an oxygen atom of one molecule and a

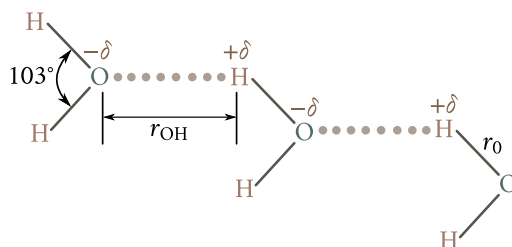


Figure 1.9: Hydrogen bond between water molecules.

hydrogen atom of another behaves as a tiny dipole with a $-\delta$ charge on the oxygen atom and a $+\delta$ charge on the hydrogen atom. The force of attraction between those charges is the cause of the hydrogen bond shown by dots in Figure 1.9. The attraction is enhanced by the small dimensions of the hydrogen atom that enable it to come close to the electronegative atom. Still, this distance $r_{OH} = 2.76 \text{ \AA}$ is much greater than the radius of the covalent bond H–O, r_0 , in the water molecule itself, which is equal to 0.96 \AA . This is quite natural since the energy of the covalent bond is about an order of magnitude higher than that of the hydrogen bond. Its value for water is $21 \times 10^3 \text{ J mol}^{-1}$ to $25 \times 10^3 \text{ J mol}^{-1}$.

The hydrogen bond is the cause of association of molecules of liquids (water, acids, spirits, etc.), which results in greater viscosity, higher boiling point, abnormal thermal expansion, etc. Water may serve best to illustrate this. Should there be no hydrogen bonds between the molecules of water, its boiling point at atmospheric pressure would be not $+100^\circ\text{C}$ but -80°C and its viscosity would be lower by almost an order of magnitude. When water is heated above 0°C , the hydrogen bond is destroyed. Since the hydrogen bond is responsible for the loose structure of associated complexes, in which the water molecules are rather far apart (2.76 \AA), the destruction of such a loose structure should result in an increase in the density of water. On the other hand, an increase in the temperature of water and a corresponding increase in the intensity of thermal motion of its molecules should lead to thermal expansion and a decrease in its density. Experiment shows that in the temperature range 0°C to 4°C the first factor—the increase in density due to the disruption of the hydrogen bonds—is the prevalent one. Because of that within this range the density of water rises upon heating. Above 4°C the other factor—thermal expansion—prevails. This is why when water is heated above 4°C its density decreases, as is the case with other (normal) liquids.

§ 6. Comparison between bonds of various kinds

The van der Waals bond is the most universal one. It exists in all cases without exception. At the same time this is the weakest, having an energy of the order of 10^4 J mol^{-1} . Ideally, it operates between neutral atoms, or molecules, with closed inner electron shells. Specifically, the van der Waals forces are responsible for the existence of the liquid and solid states of inert gases, hydrogen, oxygen, nitrogen and many other organic and inorganic compounds. They also, as we will see later, form bonds in many of the molecular valence crystals. Because of low energy values of the van der Waals bond all structures based on it are unstable, volatile and have low melting points.

The ionic bond is a typical chemical bond very frequent among the inorganic compounds such as metal-halide compounds, metallic oxides, sulfides and other polar compounds. The ionic bond is also a feature of numerous intermetallic compounds (carbides, nitrides, selenides, etc.). The energy of the ionic bond is much higher than that of the van der Waals bond and may be as high as 10^6 J mol^{-1} . Because of that solids based on the ionic bond have high sublimation heat values and high melting points.

The covalent bond is extremely widespread among organic compounds, but is also present in inorganic compounds, in some metals and in numerous intermetallic compounds. This bond is responsible for the existence of valence crystals of the diamond, germanium and other types, as will be discussed below. The energy of the covalent bond is also high ($\sim 10^6 \text{ J mol}^{-1}$), which stems from the fact that the solids with this type of bond have high melting points and high heats of sublimation.

The metallic bond formed as a result of the collectivization of the valence electrons is characteristic of typical metals and numerous intermetallic compounds. The order of magnitude of the energy of this type of bond is comparable to that of the energy of covalent bond.

Lastly, the hydrogen bond, although relatively weak, still plays an important part in nature.

It should be pointed out that in real solids no types of bonds discussed above ever exist purely by themselves. Practically, there is always a superposition of two types of bonds or more. One of them, as a rule, plays a dominant part in determining the structure and the properties of the solid.

§ 7. Forces of repulsion

For the formation of a stable system of interacting atoms or molecules, together with forces of attraction there should be forces of repulsion, which would prevent the complete merging of the particles.

The origin of the forces of repulsion is first of all the interaction of the nuclei each of which carries a considerable positive charge. The energy of this interaction, U'_{rep} depends on the distance between the nuclei and on the degree of screening by their internal electron shells.

The following expression for U'_{rep} may be obtained from quantum mechanical calculations:

$$U'_{\text{rep}} \propto e^{-r/a} \quad (1.14)$$

where r is the distance between the nuclei, and $a = 0.529 \text{ \AA}$ the Bohr radius.

This dependence of U'_{rep} on r determines the nature of the forces of repulsion: they attain enormous values at short distances and fall off abruptly as r increases. For instance, when the distance between a proton and a hydrogen atom decreases from $r = 2a$ to $r = a/2$ (4 times), the repulsive energy increases almost 300-fold.

The repulsive forces due to the interaction of the nuclei play a dominant role when light atoms, whose nuclei are rather poorly screened by the electron shells, are brought together. In all other cases the dominant part is played by repulsion due to the overlapping of closed electron shells of the atoms being brought together. Consider by way of an example the interaction between a chlorine ion with a closed 3p shell and a sodium ion with a closed 2p shell. When the ions are brought together to a distance at which the 3p and 2p shells overlap, the number of electrons in each of them begins to exceed that which is compatible with the Pauli exclusion principle. Because of that some of the electrons are forced to go to higher energy levels (for instance, 3d or 4s). This results in an increase in the system's energy and, consequently, in the appearance of forces of repulsion. Quantum mechanical calculations show the energy of such repulsion to have an exponential dependence on the distance, as well:

$$U''_{\text{rep}} \propto e^{-r/\rho} \quad (1.15)$$

where ρ is a constant usually obtained experimentally.

Often the repulsion energy is expressed in the form (1.7). This expression gives a less steep decline of U_{rep} with the increase in r and is less consistent with experiment than (1.14) or (1.15) but is nevertheless widely used by researchers.

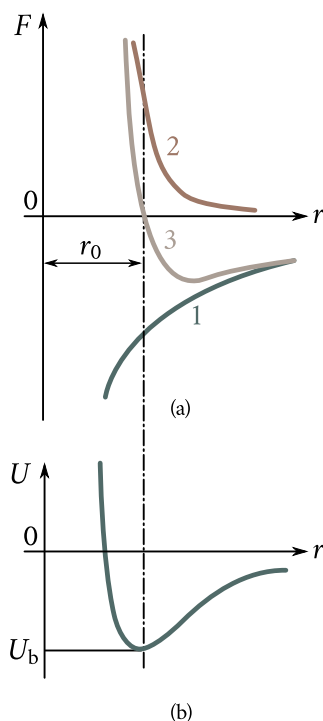


Figure 1.10: Interaction between approaching particles: (a) — interaction forces, 1 — force of attraction, 2 — repulsive force, 3 — resultant force; (b) — interaction energy.

§ 8. Crystal lattice

No matter what the origin of forces appearing when particles are brought together is, their general nature is the same [Figure 1.10(a)]: at comparatively large distances forces of attraction F_{att} increase rapidly as the distance between the particles decreases (curve 1); at small distances forces of repulsion F_{rep} come into being and with a further decrease in r they increase much more rapidly than F_{att} (curve 2). At the distance $r = r_0$ the repulsive forces counterbalance the forces of attraction, the resultant force of interaction F vanishes (curve 3), and the energy attains its minimum value U_b [Figure 1.10(b)]. Because of this the particles brought together to a distance r_0 are in a state of stable equilibrium. By the same reason the free particles should arrange themselves in a strict order at a distance r_0 from one another thus forming a body with a regular internal structure—a crystal. Such a structure will remain stable until the absolute value of the bond energy remains greater than the energy of thermal motion of the particles. The particles constituting the crystal cannot freely leave their equilibrium sites because this involves an increase in their

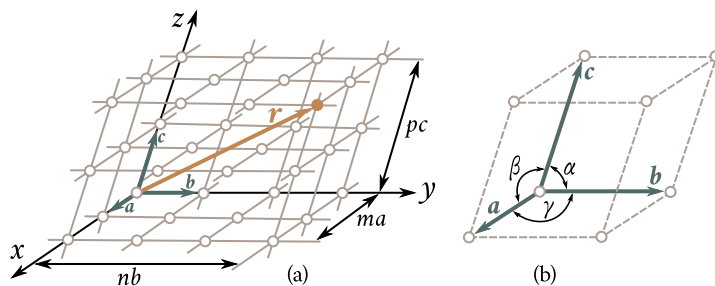


Figure 1.11: Crystal lattice: (a) — Bravais lattice, (b) — unit cell of Bravais lattice.

energy and leads to the appearance of forces tending to return them to their equilibrium sites. One may say that the particles are fixed in their equilibrium sites. The only form of motion allowed to them are random vibrations around their equilibrium sites.

To describe the regular internal structure of crystals one may conveniently use the concept of the crystal lattice. There are lattices of two types—the translational Bravais lattice and the lattice with a basis.

Bravais lattice. From the geometrical point of view a regular periodic arrangement of particles may be described with the aid of a translation. Figure 1.11(a) shows a lattice obtained with the aid of translation of a particle along the three axes: OX over the sections $a, 2a, 3a, \dots, ma, \dots$; OY over the sections $b, 2b, 3b, \dots, nb, \dots$; OZ over the sections $c, 2c, 3c, \dots, pc, \dots$ (m, n, p are integers). The position of any particle in this lattice is described by the vector

$$\mathbf{r} = m\mathbf{a} + n\mathbf{b} + p\mathbf{c}. \quad (1.16)$$

The vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are termed the *translation vectors* and their numerical values the *translation periods*.

A lattice built with the aid of translation of any site along the three directions is termed a *Bravais lattice*. The smallest parallelepiped built on the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ is termed the *unit cell* of the crystal (Figure 1.11(b)). The shape and the volume of all the unit cells comprising the lattice are identical. All cell apexes are occupied by identical atoms or groups of atoms and are therefore equivalent. They are termed *lattice sites*.

To describe a unit cell, six quantities should generally be stated: three edges of the cell (a, b, c) and three angles between them (α, β, γ). Those quantities are termed the *parameters* of the unit cell. Often the sections a, b, c are used as units of length in lattices instead of the metre. They are termed *axial units*.

Unit cells with particles only at the vertices are known as *primitive cells*. There is only one particle to each such cell.

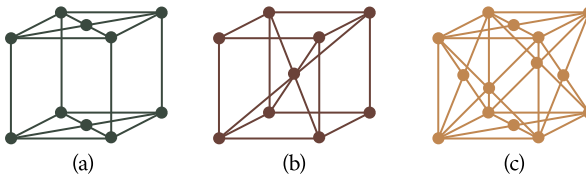


Figure 1.12: Typical crystal structures: (a) — base-centered (BaseC); (b) — body-centered (BC); (c) — face-centered (FC).

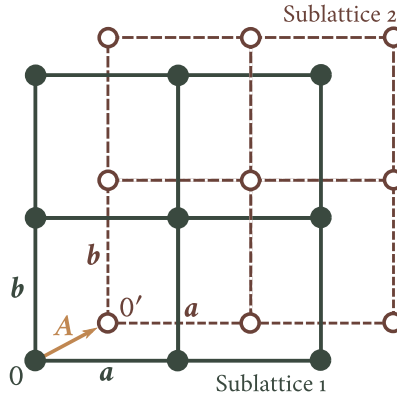


Figure 1.13: Two-dimensional lattice with a basis: \mathbf{A} — basis vector.

In some cases to express the lattice symmetry more fully the unit cells are built so that they contain particles not only in their apexes but at other points as well. Such cells are termed *complex cells*. The most widespread types of such cells are (Figure 1.12) the body-centered (BC), the face-centered (FC), and the base-centered (BaseC) cells. It may be shown that such cells may easily be reduced to primitive cells. Because of that they are Bravais-type cells.

A lattice with a basis. Not every type of lattice may be obtained by translation of a single site. Figure 1.13 shows a two-dimensional lattice with a general-type basis. It may easily be seen that it is impossible to describe the unit cell of such a lattice in terms of a single-site unit cell. Such a lattice may be imagined as consisting of two Bravais lattices, 1 and 2, each determined by the basis vectors \mathbf{a} and \mathbf{b} and inserted into each other. The relative displacement of the lattices is described by an additional basis vector \mathbf{A} . The number of such basis vectors may be arbitrary.

The lattice of this type is termed the *lattice with a basis*. It may be built with the aid of the same translations as can be used to build any of the Bravais lattices that make it up. However, in this case we shall have to translate not one site but several sites—the *basis*, defined by the totality of the basis vectors. Thus, the two-dimensional lattice shown in Figure 1.13 may be obtained by a translation of the

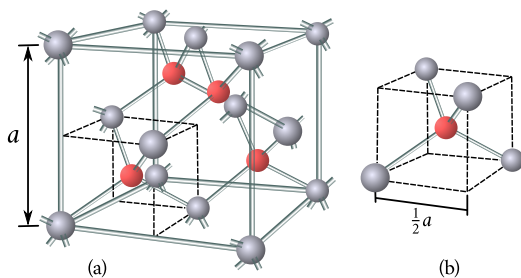


Figure 1.14: Diamond lattice: (a) — spatial arrangement of atoms in the lattice; (b) — tetrahedral pattern of atoms in the lattice.

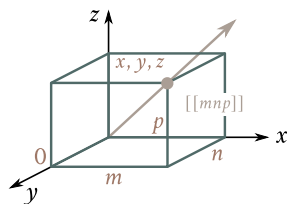


Figure 1.15: Indices of a crystal lattice site.

basis made up of two sites: 0 and 0'. An example of a three-dimensional lattice with a basis is the diamond lattice shown in Figure 1.14(a). It may be obtained by inserting one FCC (face-centered cubic) lattice into another FCC lattice displaced along the space diagonal by one-fourth of its length. Figure 1.14(b) shows a tetrahedron designated by a dotted line in Figure 1.14(a). It may be seen from Figure 1.14(b) that in the diamond lattice every atom is surrounded by four nearest neighbours in the apexes of the tetrahedron whose edge is $a/2$.

§ 9. Notation to describe sites, directions, and planes in a crystal

Let us mention briefly the notation conventionally used to describe sites, directions and planes in a lattice, the *Miller indices*.

Site indices. The position of any lattice site relative to the chosen origin of coordinates is defined by three of its coordinates (Figure 1.15): x, y, z . These coordinates may be expressed in the following form:

$$x = ma, \quad y = nb, \quad z = pc$$

where a, b, c are the lattice parameters, and m, n, p are integers.

Should we use lattice parameters as units of length along the respective axes we would obtain the lattice coordinates simply in the form of numbers m, n, p . These numbers are termed *site indices* and are written in the form $[[mnp]]$. For a negative index the minus sign is written above the index. For instance, for a site with coordinates $x = -2a, y = -lb, z = 3c$ the indices are written in the form $[\bar{2}\bar{1}3]$.

Indices of direction. To describe a direction in a crystal a straight line passing through the origin is chosen. The position of this is uniquely defined by the indices of the first site through which it passes (Figure 1.15). Therefore the indices of the

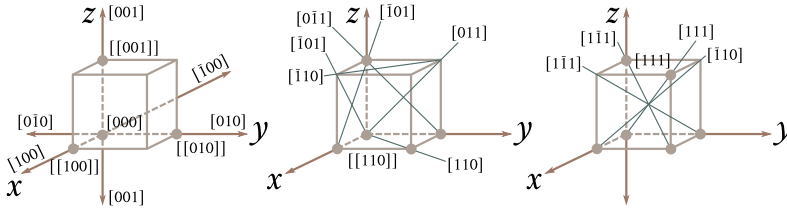


Figure 1.16: Indices of principal directions in a cubic crystal.

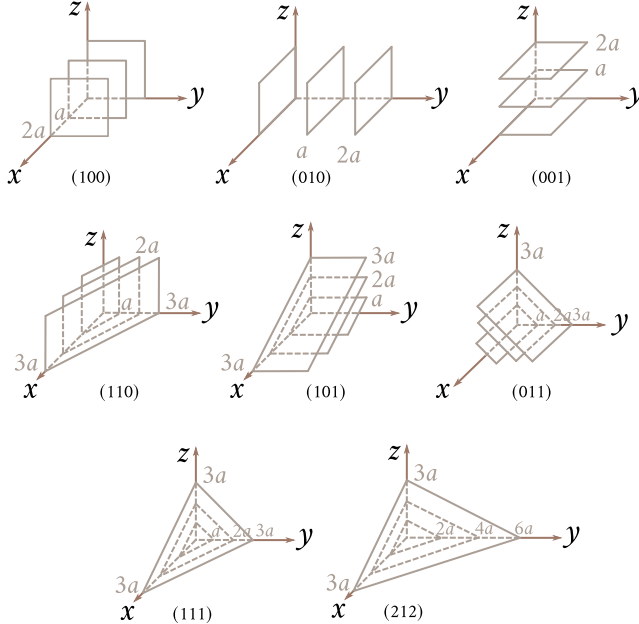


Figure 1.17: Indices of principal planes in a cubic crystal.

site are at the same time the indices of the direction. The usual notation for a direction is $[mnp]$. The indices of direction are, by definition, the three smallest integers that describe the position of the site nearest to the origin which lies on the given direction. For instance, the indices of the direction that passes through the origin and the site $[[435]]$ are $[435]$. Figure 1.16 shows the principal directions (crystallographic orientations) in a cubic crystal and their notation.

Plane indices. The position of a plane is defined by the choice of three sections A, B, C it cuts off when it intersects with the three coordinate axes. The procedure of finding the indices of such a plane is as follows.

The sections ABC are expressed in axial units and the reciprocal quantities are written as $1/A, 1/B, 1/C$. A common denominator is found for all the three

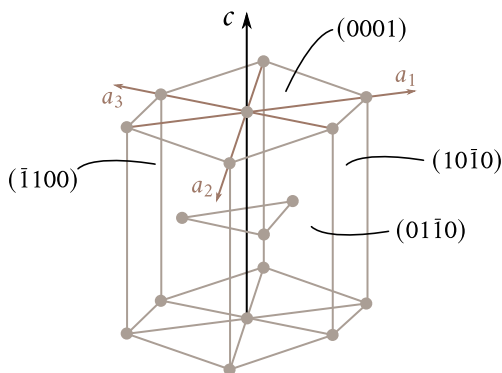


Figure 1.18: Indices of principal planes in a hexagonal crystal.

fractions. Suppose it is D . Then the integers $h = D/A$, $k = D/B$, $l = D/C$ will be the plane indices. They are written in the form (hkl) .

Determine, for example, the indices of a plane that cuts off the sections $A = 1/2$, $B = 2$, $C = 1/3$ on the x , y , z axes respectively. The ratios $1/A \div 1/B \div 1/C = 1/(1/2) \div 1/2 \div 1/(1/3) = 2 \div 1/2 \div 3$. The common denominator is $D = 2$. The indices of the plane are $h = 2/(1/2) = 4$, $k = 2/2 = 1$, $l = 3/(1/2) = 6$. The plane is denoted (416) . Figure 1.17 shows the principal planes of the cubic lattice.

It may easily be shown that in a cubic crystal the distances between the planes belonging to a given family may be expressed with the aid of the indices of these planes in the following way:

$$d = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \quad (1.17)$$

where a is the lattice parameter. This formula shows that the greater are the plane indices the shorter is the distance between the planes.

To denote the planes in a hexagonal crystal a four-axis coordinate system is used (Figure 1.18): three axes (a_1 , a_2 , a_3) make angles of 120° with one another and lie in the base of a hexagonal prism, the fourth axis, c , being perpendicular to the base plane. Every plane is denoted by four indices: $hkil$. The additional label i occupies the third place and is calculated with the aid of h and k : $i = -(h + k)$. The base plane parallel to the axes a_1 , a_2 , a_3 has the indices (0001) . The planes parallel to the lateral faces of the prism have indices of the (1010) type. There are three such planes (not parallel to one another). They are termed *first-order planes*.

§ 10. Classification of solids based on the nature of bonds

The nature of the crystal structure is primarily dictated by the nature of bonding forces acting between the structural particles (atoms, ions, molecules) which make up the solid. In accordance with the five existing types of bonds there are five principal types of crystal lattices: *ionic*, or *coordination*, *lattices*, with the ionic bond playing the main part; *molecular lattices*, with the van der Waals forces mainly responsible for the bonding; *atomic lattices*, with bonds of a distinctly covalent type; *metallic lattices*, with characteristic metallic bonds; and lattices with the *hydrogen bond*.

Let us analyze from this viewpoint the crystal structure of chemical elements and of typical chemical compounds (see Appendix Sec. A.4, Table A.1).

The chemical elements may be roughly divided into four classes according to the type of crystal structure. The analysis may best be started with Class IV.

Class IV. This class includes all the inert gases. In the process of compression and crystallization of these gases only comparatively weak van der Waals forces act between the atoms, which have spherically symmetrical electron shells. Acted upon by these forces the symmetrical atoms draw together to form a most tightly packed face-centered cubic lattice. Every atom is surrounded by 12 nearest neighbours. The number of nearest neighbours is usually termed the *coordination number* of the lattice.

Class III. The Class III includes silicon and carbon from the short periods of the Mendeleev periodic table, germanium and tin from Group IVB, and all the elements from Groups VB, VIB, VIIB.

The crystallization of the elements of those classes proceeds in conformity with the $(8 - N)$ -rule: every atom in the lattice is surrounded by $8 - N$ nearest neighbours, N being the number of the group to which the element belongs. Thus diamond, silicon, germanium and gray tin are elements of Group IV of the Periodic Table. Therefore the coordination number of their lattices should be $8 - 4 = 4$. They all do have a tetrahedral lattice in which every atom is surrounded by 4 nearest neighbours, as is shown in Figure 1.19(a). Phosphorus, arsenic, antimony and bismuth belong to Group V. Their coordination number is $8 - 5 = 3$. Every atom has 3 nearest neighbours lying in a plane, as shown in Figure 1.19(b). Their lattice has a laminate structure, with the atomic layers bonded by van der Waals forces.

Selenium and tellurium belong to Group VI and have a coordination number 2. Their atoms form long spiral-shaped chains with each atom having two nearest neighbours [Figure 1.19(c)]. The chains are bonded by van der Waals forces. Lastly, iodine belongs to Group VII [Figure 1.19(d)] and has a coordination number 1. The atoms in the iodine lattice are arranged in pairs bonded by van der Waals forces.

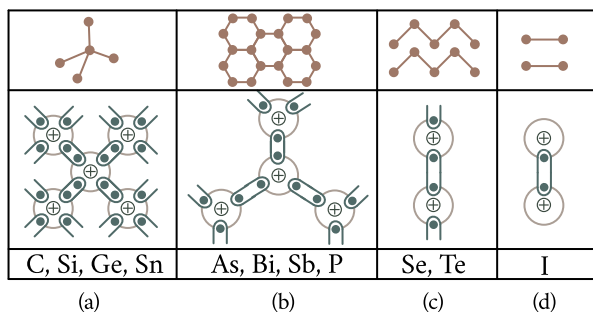


Figure 1.19: Crystal structure of chemical elements crystallizing in accordance with the $(8 - N)$ -rule: (a) — elements of Group IVB; (b) — elements of Group VB; (c) — elements of Group VIB, (d) — elements of Group VIIB.

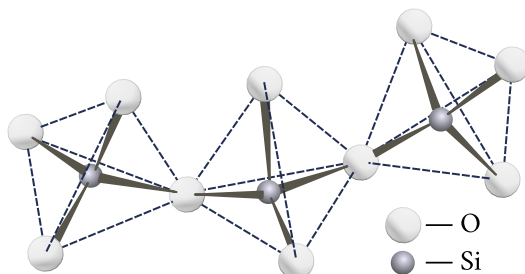


Figure 1.20: Structure of quartz SiO_2 crystal.

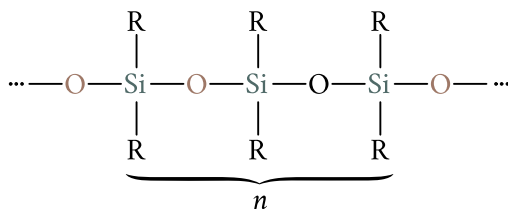
This explains the high volatility of iodine.

Such nature of the crystal structure of chemical elements whose crystallization conforms to the $(8 - N)$ -rule is quite understandable. The atoms of Group IV elements have 4 electrons in their outer shell. They lack 4 additional electrons to make up a stable 8-electron configuration. They make up the deficit by exchanging electrons with 4 nearest neighbours, as shown in Figure 1.19(a), forming a strong covalent bond with each of them. Accordingly, every atom in the crystal lattice of those elements has 4 nearest neighbours. In the same way the electron shells of Groups V, VI, VII are completed to contain 8 electrons.

Many chemical compounds crystallize in crystals with covalent bonds. Quartz SiO_2 may serve as a typical example. In the quartz crystal every silicon atom is surrounded by a tetrahedron of oxygen atoms (Figure 1.20) bonded to the silicon atom by covalent bonds. Every oxygen atom is bonded to two silicon atoms thereby joining two tetrahedrons. In this way a three-dimensional net of Si–O–Si bonds is formed, and the hardness and the melting point of the resulting crystal are high.

It may be of interest to note that the Si–O–Si bonds may be arranged into a

one-dimensional chain. Such compounds described by the common formula



where R is an arbitrary organic group, are termed *silicones*. The number n in a chain may be as high as several million. The chains may be joined together with the aid of the lateral groups R . In this way new materials, silicone resins, are formed. Because of the high strength of the Si–O–Si bonds and of the high flexibility of silicone chains such resins retain their properties at much lower and much higher temperatures than natural rubbers. This fact enables them to be used for thermal shielding of space ships and aircraft, as well as in extreme arctic conditions.

Class I. This is the most populated class which contains metals. Since metallic lattices are made up not of atoms but of ions, having the spherical symmetry of the atoms of inert gases, it is to be expected that metals too should crystallize into the same tightly packed lattices as the inert gases. Indeed, the following three types of crystal lattices are characteristic for metals: the face-centered cubic lattice with the coordination number 12 (see Figure 1.12), the hexagonal close-packed (HCP) lattice with the coordination number 12 (see Figure 1.18) and the body-centered cubic lattice with the coordination number 8 (see Figure 1.12). The latter is the least closely packed metal lattice.

Class II. The chemical elements belonging to Class II are in a sense intermediate between metals and the Class III elements, which crystallize in conformity with the $(8 - N)$ -rule. For instance, the Group IIB elements Zn, Cd and Hg are metals and one would expect them to have a typically metallic lattice with a high coordination number. Actually, Zn and Cd crystals are a special modification of the compact hexagonal lattice in which every atom has 6 nearest neighbours instead of 12, as required by the $(8 - N)$ -rule. These atoms occupy the base plane. In the case of mercury the $(8 - N)$ -rule is observed even more strictly: its crystal structure is a simple rhombohedral in which every atom is surrounded by 6 nearest neighbours. Boron—an element of Group IIIB—has a lattice that may be described as a deformed lattice with 5 nearest neighbours. This too agrees with the $(8 - N)$ -rule.

The ionic bond, as was stated above, plays one of the main parts in the world of inorganic compounds, in particular, in numerous ionic crystals typically represented by the rock salt crystal NaCl (Figure 1.21). In such crystals it is impossible to pick out single molecules. The crystal should be regarded as a closely packed

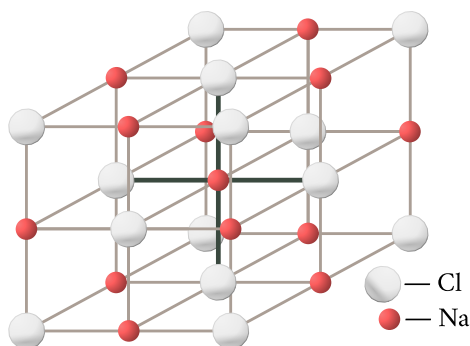


Figure 1.21: Structure of rock salt NaCl crystal.

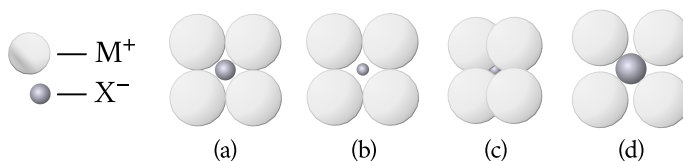


Figure 1.22: Effect of relative dimensions of ions on their packing in the lattice.

system of positive and negative ions whose positions alternate so that the electrostatic attraction between the nearest neighbours would be at its maximum. With the most favourable relation between the radii of the positive (M^+) and the negative (X^-) ions which exists in the NaCl crystal the ions “touch” one another [Figure 1.22(a)] and the closest possible packing is achieved, in which every ion is surrounded by 6 nearest neighbours of the opposite charge. When the ratio of the radii of the ions M^+ and is less favourable [Figure 1.22(b,c)] crystal structures with other coordination numbers, 4 or 8, are formed.

Ionic compounds of the MX_2 type, such as $CaCl_2$ and Na_2O , have still more complex lattices. But the principle upon which they are built remains the same: the ions are packed so as to be surrounded by ions of the opposite sign in accordance with the formula of the compound and the ratio of their radii.

Finally, let us consider crystals featuring the hydrogen bond. A typical representative of such crystals is ice. Figure 1.23(a) shows a two-dimensional diagram of the arrangement of water molecules in an ice crystal: each molecule is surrounded by four nearest neighbours a distance $r_{OH} = 2.76 \text{ \AA}$ away from it with whom it forms hydrogen bonds. In space the molecules occupy the vertices of a regular tetrahedron [Figure 1.23(b)]. The combination of such tetrahedra forms the regular crystal structure of ice [Figure 1.23(c)]. The structure is very loose and this is the cause of the abnormally low density of ice. Upon melting, some ($\sim 15\%$) of the hy-

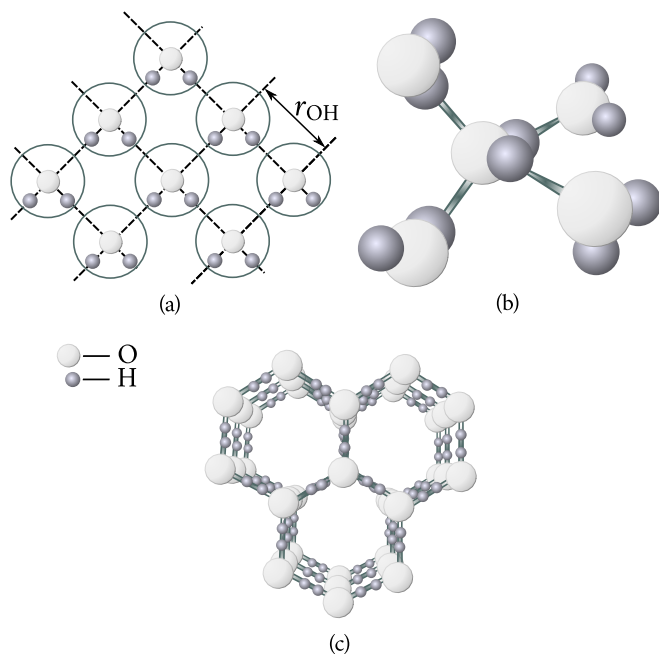


Figure 1.23: Crystals with hydrogen bond: (a) — plane diagram of arrangement of water molecules in an ice crystal; (b) — spatial arrangement of water molecules in an ice crystal; (c) — crystal structure of ice.

drogen bonds are disrupted and the packing density of the water molecules rises somewhat with the resultant rise in the density of water: the density of ice at 0°C is 916.8 kg m^{-3} , and the density of water at this temperature is 999.87 kg m^{-3} .

It may be of interest to note that if there was no hydrogen bond the melting point of ice would be -100°C instead of 0°C .

It should finally be stressed again that the hydrogen bond plays an extremely important part in vital biological compounds: protein molecules owe their helical shape exclusively to the hydrogen bond; the same type of bonds holds together the double helixes in the DNA. “It is no exaggeration to claim that life on our planet would have assumed radically different forms—if any at all—were hydrogen bonding not present in water and in the proteins and nucleic acids that compose living cells and that transmit hereditary traits”².

Table A.2 of Appendix A.4 shows the general classification of solids. The upper left corner contains typical metals with collectivized electrons (silver, copper) and

²G. C. Pimentel and R. D. Spratley, *Chemical Bonding Clarified Through Quantum Mechanics*, Holden-Day, San Francisco (1969), p. 261.

the upper right corner typical valence crystals with distinctly localized electron bonds. The extreme righthand part contains crystals with van der Waals bonds. Such elements as silicon and germanium occupy an intermediate position between the metals and the valence crystals. At absolute zero they are typical valence crystals; however, as temperature rises the valence bond is gradually disrupted and they begin to exhibit metallic properties. Such solids as sulphur, phosphorus, and selenium occupy an intermediate position between the valence crystals and crystals with the van der Waals bond.

The lower left corner of the diagram contains alloys of the NiCu type with the characteristic metallic bond and the lower right corner —ionic crystals (sodium chloride). Intermediate positions between them are occupied by numerous intermetallic compounds of the Mg_3Sb_2 type featuring the ionic bond (Mg_3Sb_2 corresponds to a $\text{Mg}^{+2}-\text{Sb}^{-3}$ compound). Intermediate position between the ionic and the valence crystals is occupied by such compounds as SiO_2 and SiC , with bonds of partially ionic nature made possible because of electron displacement. Compounds of the FeS and the TiO_2 (titanium dioxide) type occupy an intermediate position between ionic crystals and crystals with the van der Waals bond.

There are a great many crystals in which ionic or covalent bonds act in atomic planes while the bonds between the planes are of the van der Waals type.

§ 11. Polymorphism

Some solids have two or more crystal structures each of which is stable in an appropriate range of temperatures and pressures. Such structures are termed *polymorphic modifications*, or *polymorphs*, and the transition from one modification to another, *polymorphic transformation*.

It is the practice to denote polymorphic modifications by Greek letters: the modification stable at normal and lower temperatures is denoted by α ; modifications stable at higher temperatures are denoted by the letters β , γ , δ , etc. The polymorphism of tin may serve as a classical example. Below 13.3°C the stable modification of tin is α -Sn, which has a tetragonal cubic lattice of the diamond type. This is the so-called gray tin. It is brittle and may easily be ground to powder. Above 13.3°C α -Sn transforms into β -Sn, which has a body-centered tetragonal lattice. This is the familiar white metallic tin, a rather ductile metal. The transformation from β -Sn to α -Sn is accompanied by a considerable increase in specific volume (by about 25%). Long ago when many things were made of tin, the perplexing phenomenon of growing bulges on them and their subsequent destruction following excessive cooling was attributed to a mysterious metal disease, the “tin plague”.

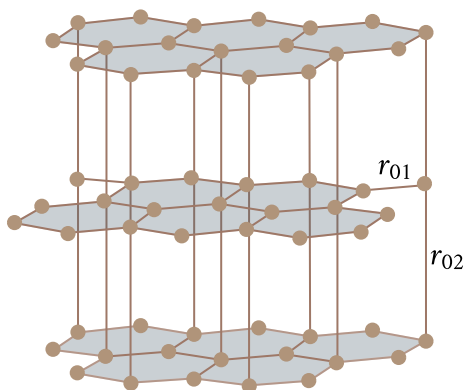


Figure 1.24: Crystal structure of graphite.

Many other chemical elements also exhibit polymorphism: carbon, iron, nickel, cobalt, tungsten, titanium, boron, berillium, and others, as well as many chemical compounds and alloys.

An interesting and a practically important case of polymorphism is that of carbon, which exists in the forms of diamond and graphite.

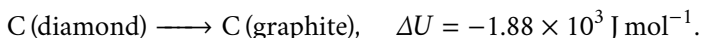
This case deserves a more detailed study. In the diamond lattice every atom is surrounded by 4 nearest neighbours occupying the vertices of a tetrahedron (see Figure 1.14) with whom it is bonded by strong covalent forces. The length of the bond is 1.544 \AA and the energy per bond is about $3.5 \times 10^5 \text{ J mol}^{-1}$.

The graphite lattice is of the pattern characteristic of the Group VB element lattices: the carbon atoms form two-dimensional layers in which every atom is surrounded by 3 nearest neighbours with whom it is bonded by covalent bonds (Figure 1.24). The length of the bond is $r_{01} = 1.42 \text{ \AA}$, that is, less than in the diamond lattice, therefore the former is stronger. The distance between the layers is much greater than the length of the C–C bond, being equal to $r_{02} = 3.6 \text{ \AA}$. Only weak van der Waals forces can act at such great distances and the layers are held together by them. The energy of this bond is $4 \times 10^3 \text{ J mol}^{-1}$ to $8 \times 10^3 \text{ J mol}^{-1}$.

Such a great difference in the nature of the bonding forces in the diamond and graphite structures should evidently manifest itself in a great difference in their properties, which is actually the case. Graphite slides easily along the planes held together by weak van der Waals forces. Therefore it is used to advantage in making “lead” pencils and as dry lubricant. The electrons in diamond are held securely between the atoms forming bonds. Light of the visible part of the spectrum is unable to knock out such electrons and therefore is not absorbed in diamond. Because of this diamond is an ideal transparent crystal unable to conduct electric current (a di-

electric). In graphite one of the four valence electrons of the carbon atom is actually collectivized by the atoms forming the layer. Such electrons can easily be moved by the action of an external electric field, making graphite a two-dimensional conductor. The presence of mobile electrons explains light absorption (the gray colour of graphite) and its characteristic metallic glitter.

In normal conditions graphite is a somewhat more stable modification than diamond, although the difference in the energies of those modifications is quite small—of the order of $2 \times 10^3 \text{ J mol}^{-1}$:



Still, such a difference is enough to bring about a sufficiently rapid transformation of diamond into graphite when heated above 1000°C in the absence of air.

The density of diamond is greater than that of graphite (3500 and 2250 kg m^{-3} , respectively), which is due to a loose packing of the atomic layers in graphite. Therefore at greater pressures diamond becomes more stable and graphite less stable. At sufficiently high pressures diamond becomes more stable than graphite. In such conditions by raising the temperature to increase the mobility of the carbon atoms we may bring about the transformation of graphite into diamond.

The conditions for such transformation to proceed at a practical rate were calculated by the Soviet physicist O. I. Leipunskii. He writes: “Firstly, graphite should be heated to at least 2000°C for the carbon atoms to be able to move from place to place. Secondly, it must be subjected to very high pressure, not less than 60000 atm ”³. These conditions were first achieved by the scientists of the General Electric Research and Development Center, who in 1954 succeeded in producing the first synthetic diamonds in the form of dark unsightly crystals, the largest being 1.5 mm long. Subsequently, the synthesis of diamonds was mastered in Sweden, the Netherlands, and Japan.

In the Soviet Union the production of synthetic diamonds on a commercial scale began in 1961. The pressure in the process is of the order of 100000 atm and the temperature about 2000°C . Synthetic diamonds produced by this process are harder and stronger than natural diamonds and their industrial use is about 40% more efficient than that of natural ones.

Another material of extreme hardness had been synthesized in a process similar to that of the diamond—the cubic boron nitride BN, which became known as *borazon*. It is harder than diamond and may be heated up to 2000°C in atmospheric conditions. In its hexagonal modification boron nitride is similar to graphite—a white powder oily to the touch.

From the theoretical point of view all solids should exhibit polymorphism pro-

³Leipunskii, O. I.: Quoted from I. I. Shafranovskii, *Diamonds*, “Nauka”, Moscow (1964).

vided the range of their stability is not limited by the processes of melting and sublimation. The existence of polymorphism is a direct consequence of the variation of the strength and the nature of the bonds in the crystal lattice caused by the changes in intensity of atomic motion and in the distance between them as a result of heating or of application of external pressure to the crystal. Close to absolute zero the stable structure should be that with the strongest bonds possible for the given atomic ensemble. In the case of tin, which belongs to Group IV of the Mendeleev periodic table, such structure is the diamond structure, in which every atom is bonded to 4 nearest neighbours by strong covalent bonds. However, as the temperature is raised, those bonds, because of their strict directionality and rigidity, are easily destroyed by thermal motion, and already above 13.3 °C the flexible metallic structure formed with the aid of collectivized electrons becomes more favourable. This bond has its own stable crystal structure, the tetragonal body-centred lattice.

The transition from one modification to another is accompanied by the liberation or absorption of latent heat of transformation and is therefore a phase transition of the first kind. Such a transition involves the transformation of the crystal lattice and this fact together with a low mobility of atoms in solids makes possible a practically infinitely long existence of a modification thermodynamically unstable in particular conditions. Diamond which can exist ages without turning into graphite—the stable modification in normal conditions—is a striking example of this point.

Polymorphism is very important for practical purposes. Heat treatment of steels to obtain various properties, the production of stainless steels, the treatment of various alloys to obtain the necessary properties are all to a large extent based on the use of polymorphism.

§ 12. Imperfections and defects of the crystal lattice

Mosaic structure. Numerous data obtained in the study of the structure of real crystals point to the fact that their internal structure is essentially not the same as that of an ideal crystal. To begin with, real crystals have a *mosaic structure*: they are made up of regular blocks which are only approximately parallel to one another. The dimensions of the blocks vary from 10^{-6} m to 10^{-8} m and the angles between them from several seconds to tens of minutes. Because of the different orientation of adjacent blocks there is a transition layer between them in which the lattice changes its orientation gradually from that of the first block to that of the second. Therefore in this layer the lattice is deformed as compared with that of an ideal crystal.

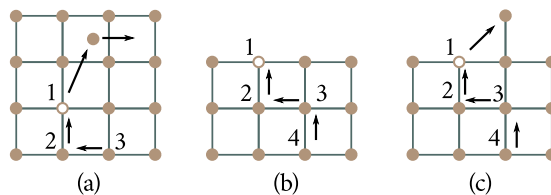


Figure 1.25: Crystal lattice defects: (a) — Frenkel defects; (b), (c) — Schottky defects.

Lattice deformations are even greater near the grain boundaries in a polycrystal, since the orientation of adjacent grains may differ by as much as tens of degrees. The grain and block boundaries carry excess free energy, which increases the rate of chemical reactions, of polymorphic transformations, of diffusion, etc. They also serve as effective carrier scattering centres responsible for a considerable part of the solid's (metal or semiconductor) electrical resistance.

Frenkel defects. The distribution of energy among the atoms of a solid is very nonuniform, as is the case with the molecules of a gas or liquid. At any temperature there are atoms in the crystal whose energy is many times greater or less than the average energy corresponding to the law of equipartition of energy. The atoms that at a given instant of time have enough energy can not only move a considerable distance away from their position of equilibrium, but can also surmount the potential barrier set up by the neighbouring atoms and move over to new neighbours, to a new cell. Such atoms acquire the capability, so to say, to “evaporate” from their lattice sites and to “condense” in its internal cavities, or interstitials [Figure 1.25(a)]. This process results in the creation of a vacant site (a *vacancy*) and of an atom in the interstitial position (a *displaced atom*). Such lattice defects are termed *Frenkel defects*.

Calculation shows the equilibrium concentration of interstitial atoms n_F at a given temperature to be

$$n_F = AN \exp\left(-\frac{E_F}{k_B T}\right) \quad (1.18)$$

where E_F is the formation energy of the interstitial whose order of magnitude is several electron volts, N is the number of sites in the given volume, A is an integer (usually about 1) indicating the number of identical interstitial positions per one lattice atom, and k_B is the Boltzmann constant.

Both the interstitial atoms and the vacancies do not remain localized in one place but diffuse through the lattice. The diffusion of a displaced atom proceeds by the motion of this atom from one interstitial position to another, and the diffusion of a vacancy by a relay process in which the vacancy is filled by neighbouring atoms [Figure 1.25(a)]: when atom 2 moves into vacancy 1 the vacancy moves over to site

2, when atom 3 moves to the now vacant site 2 the vacancy moves to site 3, etc.

Schottky defects. In addition to internal evaporation there is also a possibility of a partial or even complete evaporation of an atom from the surface of a crystal. Complete evaporation means that the atom leaves the crystal surface and joins the vapour phase [Figure 1.25(b)]. Partial evaporation means that the atom leaves the surface layer and arranges itself on top of it [Figure 1.25(c)]. In both cases a vacancy is produced in the surface layer. But when an atom from the interior of a crystal occupies a vacancy, the latter is pulled into the crystal and diffuses there. Here there are no displaced atoms to correspond to the vacancies, since the latter are produced without the simultaneous transition of atoms to interstitial positions. Such vacancies are termed *Schottky defects*. Calculations show the equilibrium number of vacancies n_{Scn} in a crystal of N sites to be

$$n_{\text{Scn}} = N \exp \left(-\frac{E_{\text{Scn}}}{k_{\text{B}}T} \right) \quad (1.19)$$

where E_{Scn} is the energy of formation of a single vacancy. It is somewhat lower than E_{F} . For instance, for aluminium it is equal to 0.75 eV. Substituting this value into (1.19), we obtain $n_{\text{Scn}} \approx 10^{18} \text{ m}^{-3}$ at $T = 300 \text{ K}$; at $T = 923 \text{ K}$, that is, close to the melting point of aluminium ($T_{\text{m}} = 933 \text{ K}$), $n_{\text{Scn}} \approx 10^{25} \text{ m}^{-3}$. Such values are characteristic of all metals at temperatures close to their melting points.

The energy of formation of the Frenkel defects is approximately equal to the sum of formation energies of the vacancy and the interstitial atom.

The Frenkel and Schottky defects play an important part in many processes in crystals. They act as carrier scattering centres reducing their mobility. They can also act as sources of carrier production, that is, play the role of donors and acceptors (usually the latter). They can also appreciably affect optical, magnetic, mechanical, and thermodynamic properties of crystals, especially of thin semiconducting films and fine crystalline specimens (because defect concentration in them is usually much greater than in bulk specimens).

Impurities. Impurities are one of the most important and most common type of defects in the structure of real crystals. Contemporary refining methods are unable to guarantee absolute purity of materials. Even the most pure materials contain up to 10^{-9} percent of impurities, which corresponds to a concentration of about 10^{17} impurity atoms per cubic metre of the material. To illustrate this degree of purity we would like to cite an equivalent example of one grain of rye contained in about 10 tons of wheat.

The impurities contained in the crystal may, depending on their nature, be in the form of dissolved atoms or in the form of inclusions of various dimensions. In the process of dissolution the impurity atoms enter the interstitial positions

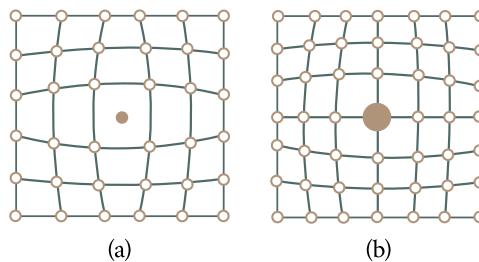


Figure 1.26: Deformations of crystal lattice of solid solutions: (a) — interstitial; (b) — substitution.

between the atoms of the crystal or substitute some of them in their sites. The solid solution of the first type is termed the *interstitial solution* [Figure 1.26(a)] and that of the second the *substitution solution* [Figure 1.26(b)]. Because of a difference in the physical nature and the dimensions of the impurity atoms from the atoms of the crystal, their presence results in the deformation of the crystal lattice.

Impurities may appreciably affect chemical, optical, magnetic, and mechanical properties of solids. They are effective carrier scattering centres, being the cause of electrical resistance that does not vanish at absolute zero temperature. In semiconductor crystals the impurities create new energy levels leading to the appearance of impurity conductivity. Calculations show that a perfectly pure silicon should have a specific resistance of the order of $2000\ \Omega\ \text{m}$. Active impurities contained in it in a concentration of 10^{-9} percent reduce the resistivity to several units. Technically pure germanium was for a long time regarded as a metal because its resistivity was of the same order as that of metals. Only perfect refining methods that brought impurity concentration in germanium down to 10^{-7} – 10^{-8} percent made it a typical semiconductor.

Interesting results were obtained in the course of investigations into the properties of extremely pure metals. Thus thoroughly purified iron turned out to be chemically inert and immune to corrosion even in conditions of tropical humidity. Titanium, chromium, bismuth, tungsten, molybdenum, which had a reputation for brittleness, turned out to be ductile even in conditions of extreme cooling; tin purified to contain no more than 5×10^{-6} percent impurities is so soft that it bends under its own weight like dough.

Some striking results were obtained in dehydration experiments: materials dried so as to contain negligible amounts of residual moisture change their properties in a marked degree. Thus dried oxyhydrogen gas does not explode at high temperatures; carbon monoxide does not burn in oxygen; sulphuric acid does not react with alkali metals, etc. The English chemist H. B. Baker sealed 11 thoroughly

purified individual chemical compounds in glass tubes together with phosphoric anhydride (a powerful absorber of water). The tubes were opened 9 years later in conditions that precluded the admission of moisture. The results were startling: the boiling point of all the compounds rose appreciably. For instance, the boiling point of benzol turned out to be 26 °C higher than that specified in tables, that of ethyl alcohol was 60 °C higher, that of bromine was 59 °C higher, and that of mercury was almost 100 °C higher. Subsequent experiments carried out by other investigators substantiated those results. More than that, it was established that very dry materials not only change their boiling point but melting point and other properties as well.

Despite substantial progress in the field of production of ultrapure materials there is a growing demand for better purification methods and presently there will be a need for materials with impurity concentrations of no more than 10^{-10} - 10^{-12} percent. This applies in the first instance to materials used for thermonuclear fusion apparatus, to microelectronics materials, as well as to materials used in other branches of industry. Such materials are not only difficult to produce but also difficult to keep pure, especially if they have to be processed before use. To illustrate how easy it is to make a mistake while working with such materials we would like to cite a case told by the well-known German physicist Werner Heisenberg. When a target was irradiated with a flux of neutrons in a mass spectrometer, gold nuclei were detected. This effect vanished after the experimenter took off and put away his gold-rimmed eyeglasses.

Chapter 2

Mechanical Properties of Solids

The mechanical properties—strength, hardness, ductility, wear-resistance—are the most characteristic of the properties of solids. Thanks to those properties the practical use of solids as constructional, building, electrotechnical, magnetic and other materials without which the growth of economy is impossible has become so widespread. The very names of the periods of human culture reflect the names of the solids whose mechanical properties made a qualitative leap in the process of development of human society possible—the Stone Age, the Bronze Age, the Iron Age.

This chapter deals briefly with modern physical concepts concerning the mechanical properties of solids, the laws of their plastic flow and destruction, the physical nature of strength, and prospects for the development of materials with unique mechanical properties.

§ 13. Elastic and plastic deformations. Hooke's law

When a crystal is subjected to an external extension load, the distances between the atoms become greater and the atoms are displaced from their equilibrium positions in the crystal. This destroys the equilibrium between the forces of repulsion and attraction characteristic of the equilibrium state of the atoms in the lattice and results in the appearance of internal forces tending to return the atoms to their initial equilibrium positions. The value of those forces per unit cross-sectional area of the crystal is termed *stress*. Let us calculate it.

It was shown in Chapter 1 that the energy of interaction of particles 1 and 2 in a solid is a function of the distance r between them. This can be described by the curve $U(r)$ schematically shown in Figure 2.1(a). When particle 2 is displaced from its equilibrium position to a distance x , that is, when the distance between

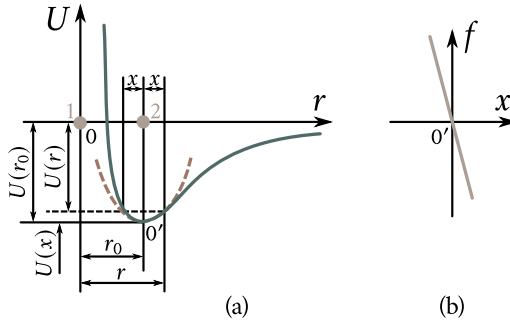


Figure 2.1: Variation of (a) interaction energy and (b) interaction force with the displacement of a particle from equilibrium position by a distance x .

the particles is increased to $r = r_0 + x$, the particle's energy grows and becomes $U(r)$. The change in energy $U(x) = U(r) - U(r_0)$ can be found if we expand $U(r)$ into a Taylor series in powers of x :

$$U(x) = \left(\frac{\partial U}{\partial r} \right)_0 x + \frac{1}{2} \left(\frac{\partial^2 U}{\partial r^2} \right)_0 x^2 + \frac{1}{6} \left(\frac{\partial^3 U}{\partial r^3} \right)_0 x^3 + \dots \quad (2.1)$$

Leaving only the quadratic term of the series and taking into account the fact that $(\partial U / \partial r)_0$ at point $0'$ is zero, we obtain

$$U(x) \approx \frac{1}{2} \left(\frac{\partial^2 U}{\partial r^2} \right)_0 x^2 = \frac{1}{2} \beta x^2 \quad (2.2)$$

where β is the rigidity of the bond.

We obtained an approximate expression for the change in energy of a particle brought about by its displacement from its equilibrium position to a distance x . This expression is an approximation because we left only the quadratic term in the expansion (2.1), neglecting higher-order terms. Graphically this dependence is expressed by a parabola shown in Figure 2.1(a) by a dotted line.

The force which appears between particles 1 and 2 when the distance between them is changed by x is equal to

$$f = - \frac{\partial U(x)}{\partial x} = -\beta x. \quad (2.3)$$

It follows from (2.3) that the force is linearly dependent on x and is directed towards the position of equilibrium, as indicated by the minus sign. It is well known that a body acted upon by such a force oscillates harmonically. Therefore such force is termed *harmonic*, the same term being applied to the approximation (2.2) (*harmonic approximation*). Figure 2.1(b) is a schematic diagram of the $f(x)$ dependence for small values of x . It is a straight line.

Now let us imagine that a tensile load F is applied to a rod with a cross-sectional

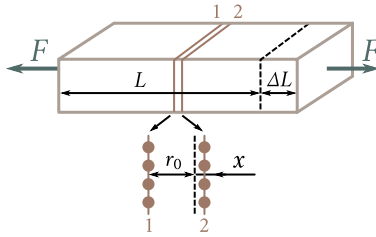


Figure 2.2: Uniaxial extension of a rod by an external force F : 1 and 2 are the schematic representation of adjacent atomic planes.

area S and a length L . This load changes the distance between the neighbouring atomic planes 1 and 2 by the amount x causing thereby an extension of the rod by ΔL (Figure 2.2). It will be counterbalanced by the internal force F_{int} equal numerically to

$$F_{\text{int}} = fN = N\beta x \quad (2.4)$$

where N is the number of particles in the atomic layer of area S .

The stresses σ which appear in the extended rod will be

$$\sigma = \frac{F_{\text{int}}}{S} = \frac{N}{S}\beta x = cx \quad (2.5)$$

where $c = N\beta/S$. Multiplying and dividing the right-hand side of (2.5) by the distance between the atomic planes, r_0 , we obtain

$$\sigma = cr_0 \frac{x}{r_0} = E\varepsilon \quad (2.6)$$

where

$$E = cr_0 = \frac{N}{S}\beta r_0 \quad (2.7)$$

is termed the *elasticity modulus*, or *Young's modulus*, and

$$\varepsilon = \frac{x}{r_0} \quad (2.8)$$

is the relative change in the lattice parameter in the direction of the external force F .

Multiplying the numerator and the denominator of (2.8) by the number of atomic layers N' contained in the sample of length L , we obtain

$$\varepsilon = \frac{xN'}{r_0N'} = \frac{\Delta L}{L}. \quad (2.9)$$

Hence, ε is the relative elongation of the sample under the action of the external tensile load.

It follows from Eq. (2.6) that as long as the harmonic approximation remains valid, that is, as long as the forces acting between the particles displaced in relation

to each other as a result of the deformation of the body remain linear functions of the displacement, the stresses σ which appear in the body will remain proportional to the relative deformation of the body:

$$\sigma = E\varepsilon.$$

The elasticity modulus E serves as the proportionality factor.

Formula (2.6) expresses the well-known *Hooke's law*. It is valid only as long as the harmonic approximation is valid, that is, only for very small relative deformations ε .

The physical meaning of the elasticity modulus is quite evident from Eq. (2.6). Putting $\varepsilon = 1$, we find that $\sigma = E$. Hence, the elasticity modulus is numerically equal to the stress which is capable of causing an elongation $\Delta L = L$ of the sample, provided Hooke's law remains valid and the sample is not destroyed. No real material except rubber can stand such deformations.

Table 2.1 shows the values of the elasticity modulus of some metallic crystals.

It follows from data presented in Table 2.1 that the elasticity modulus of solids is very large (of the order of 10^{10} Pa to 10^{11} Pa), which is an indication that the bonding forces in those bodies are very strong.

For some crystals the value of the elasticity modulus depends appreciably on the direction in which the lattice is deformed. Table 2.1 shows the values of E for directions in which it is at its minimum and at its maximum. For some crystals the ratio E_{\max}/E_{\min} may be as high as 3, pointing to a high degree of *anisotropy* of such crystals.

The elasticity modulus depends only on the nature of the atoms (molecules) making up the body and on their mutual arrangement. It can be changed only by

Table 2.1

Substance	$E(\text{GPa})$		$G(\text{GPa})$	
	maximum	minimum	maximum	minimum
Aluminium	77	64	29	25
Copper	194	68	77	31
Iron	200	135	118	60
Magnesium	514	437	184	171
Tungsten	400	400	155	155
Magnesium	126	65	497	278

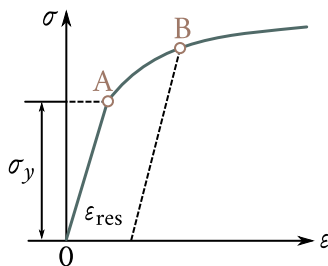


Figure 2.3: Typical extension curve of a ductile metal: σ_y — yield stress, ε_{res} — residual (plastic) deformation, OA—elastic deformation region, AB — plastic deformation region.

a substantial change in composition or internal structure of the solid. However, even in such cases the changes in E are relatively small. Thus, high concentration alloying, heat treatment, cold rolling, etc. of steel result in great improvement in its hardness and in other mechanical properties but only in negligible (up to 10%) changes in its elasticity modulus; alloying copper with zinc up to 40% leaves the elasticity modulus practically unchanged, although other properties experience a profound change.

We have discussed the tensile stress. However, all the considerations and the results obtained remain valid for other types of deformation—compression and shear—as well. In the latter case one should make use of the shear modulus G , whose values are also presented in Table 2.1.

When the external load is steadily increased, stress σ and deformation ε increase steadily too (Figure 2.3). At some stress σ_y , characteristic of the specific crystal, the crystal is either destroyed or the direct proportionality between σ and ε ceases and a residual (plastic) deformation ε_{res} sets in which remains after the external load has been removed. The first case is that of a brittle material and the second of a ductile one. The stress σ_y at which a noticeable plastic flow in the body sets in is termed the *yield stress* and OA and AB are the regions of the elastic and plastic deformations, respectively.

In brittle materials the elastic limit coincides with the tensile strength, and their destruction begins before a noticeable plastic flow sets in. In ductile metals, on the other hand, the elastic limit and the yield stress are, as a rule, much lower than the tensile strength, and these materials are destroyed only after a substantial plastic deformation has taken place.

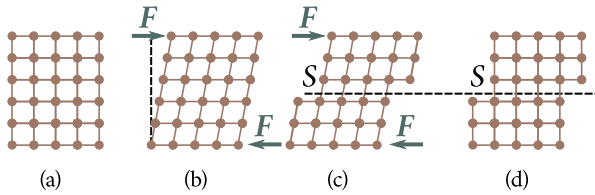


Figure 2.4: Crystal deformation by a shear force F : (a) — initial unstressed crystal; (b) — elastic deformation caused by shearing stress not exceeding elastic limit; (c) — early stages of plastic shear (slip) in the S plane caused by stress exceeding elastic limit; (d) — external force is removed, residual deformation (residual shift of one part of the lattice in relation to another) remains.

§ 14. Principal laws governing plastic flow in crystals

Residual deformations occur in all cases when the stress in ductile crystals tested for extension and compression exceeds the yield stress. However, neither extension nor compression can by themselves be the causes of such deformations. An increase in the length of the crystal can only result in an increase in the distance between the atomic planes perpendicular to the acting force. When these planes are drawn far enough apart, it may be that the forces of attraction shall no longer be able to compensate for the external load and the crystal will break. Compression can only draw the atomic planes closer together until the repulsive forces appearing between the atoms are able to counterbalance the external load. Deformation in this case is ideally elastic and cannot lead to irreversible displacement of parts of the lattice.

Plastic deformation may only be the result of shearing stresses, which are able to shift some parts of the crystal in relation to the others without destroying the bonds between them. Such displacement is termed *slipping*. It lies at the basis of the plastic flow process in crystalline materials. Figure 2.4 shows how residual deformations originate and develop in crystals [Figure 2.4(a)] acted upon by a shearing force F . As long as the elastic limit is not reached the crystal experiences elastic deformations [Figure 2.4(b)] with the tangential stresses growing in proportion to the relative shearing deformation γ (*Hooke's law*):

$$\tau = G\gamma \quad (2.10)$$

where G is the shear modulus. After the crystal is relieved from external load the atoms return to their initial positions. When the elastic limit is exceeded, one part of the crystal shifts in relation to another [Figure 2.4(c)] by one or more atomic distances along definite planes S termed *slip planes*. When the external load is withdrawn, the elastic stresses in the lattice vanish. However, one part of the crystal

remains displaced in relation to another [Figure 2.4(d)]. Such small irreversible displacements that proceed along numerous slip planes sum up to produce the residual deformation of the crystal as a whole.

The degree to which a crystal can be subjected to plastic deformations is determined, first of all, by the nature of the bonding forces acting between its structural elements.

The covalent bond with its rigorous directionality is appreciably weakened already by small relative displacements of the atoms. Shear destroys such bonds even before the atoms are able to establish them with other neighbouring atoms. On account of this the valence type crystals (such as diamond, silicon, germanium, antimony, bismuth, and arsenic) are incapable of plastic deformation. Outside the elastic deformation range they experience brittle destruction.

The metallic bond, which does not exhibit any directionality, on the other hand, remains practically unchanged as a result of relative tangential displacements of the atoms. This makes very great (some thousand atomic distances) relative displacements of some parts of the lattice possible, resulting in a high degree of ductility of crystals of this type.

The ionic bond occupies an intermediate position between the metallic and covalent bonds. It is less directional than the covalent bond but not so flexible as the metallic bond. Typical ionic crystals such as NaCl, CaF₂, and KCl are almost as brittle as the valence type crystals. At the same time silver chloride crystals are rather ductile.

Slipping takes place in crystals along definite crystallographic planes and directions, usually along the closest-packed planes and directions. This is because the closest-packed planes and directions are the strongest since the interatomic distances in them are the shortest and bonding is at its maximum. On the other hand, the distance between such planes is the greatest [see (1.17)]; on account of this the bonding between them is at its minimum. Slipping along such planes and directions results in the minimum disarrangement in atomic order and is therefore the easiest to accomplish.

The combination of the slip plane and the slip direction, which lies in it, forms the *slip system*. In the face-centered cubic lattice the slip plane coincides with the octahedral plane (111) and the slip direction with the direction of the body diagonal [111]. In hexagonal crystals the SS slip plane coincides with the base plane (0001) and the X slip direction with one of the three axes lying in the base plane (see Figure 2.5, where P is the external deforming load).

Numerous experiments have shown that the crystal begins to “slip” in the given slip system only after the shearing stress τ acting in this system reaches the critical value τ_{cr} termed the *critical shearing stress*. Table 2.2 shows the values of critical

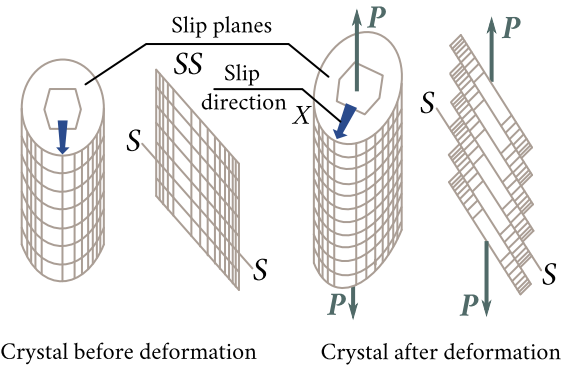


Figure 2.5: Slip planes and directions in a crystal.

shearing stresses for some pure metallic single crystals.

It follows from the data of Table 2.2 that for the most ductile single crystals the critical shearing stress does not exceed 10^6 Pa.

The critical shearing stress depends to a large extent on the prior deformation of the crystal rising with the increase in the latter. This phenomenon became known as strengthening, or *cold working*. Thus a 350% preliminary deformation of the magnesium single crystal increases τ_{cr} nearly 25 times. Even greater is the effect of cold working on the cubic crystals—aluminium, copper, nickel, etc.

The strengthening of crystals is a witness to the fact that irreversible processes involving the relative displacement of atoms and of parts of the crystal take place. This results in changes of the internal energy of the crystals. Experimental study of this phenomenon has proved that the changes in the internal energy of solids in the process of their plastic deformation do, indeed, take place. Table 2.3 shows the maximum amounts of energy that are accumulated by various metals in the

Table 2.2

Metal	Impurity content (10^{-4})	Slip plane	Slip direction	τ_{cr} (10^7 Pa)
Cadmium	0.40	(0001)	[100]	0.058
Copper	10.0	(111)	[101]	0.100
Magnesium	5.00	(0001)	[100]	0.083
Nickel	20.0	(111)	[101]	0.580
Silver	1.00	(111)	[101]	0.060
Zinc	4.00	(0001)	[100]	0.094

process of their plastic deformation.

Should this energy be transformed into heat it would suffice to heat the metal by several degrees.

Since the accumulation of energy in the crystal in the process of its plastic deformation involves irreversible displacements of the atoms and of parts of the crystal, this energy is, in effect, the energy of *residual stresses* remaining in elastically deformed parts of the crystal lattice.

Because of a higher value of internal energy in a cold worked crystal it is less thermodynamically stable than the annealed crystal. This gives rise to processes that tend to bring the crystal to the equilibrium state. Relaxation and recrystallization are two such processes.

Relaxation consists in the dissipation of internal stresses, with the atoms of the deformed parts of the lattice returning to their regular positions. This process does not involve visible changes in the crystal structure and results in a partial or complete removal of the strengthening obtained as a result of plastic deformation. Being a diffusion-controlled process relaxation proceeds at a rate that strongly depends on temperature and on the latent heat of defect formation. Metals with a low melting point (such as tin, lead, cadmium, zinc) have comparatively high self-diffusion rates already at room temperatures. Accordingly, their relaxation rates at room temperatures are quite noticeable. At the same time there is practically no relaxation at room temperature in the metals with a high melting point; but the relaxation rate rises sharply as the temperature is increased (the relaxation process goes as far in 1 minute at 315 °C as it would in a hundred years at room temperature).

Another process that also results in the disappearance of the hardening in a cold worked crystal—the *recrystallization* process—proceeds intensely at temperatures of the order of one quarter of the melting temperature of the metal (on the

Table 2.3

Metal	Q (J kg⁻¹)
Aluminium	4400
Brass	2000
Copper	2000
Iron	4800
Nickel	3120

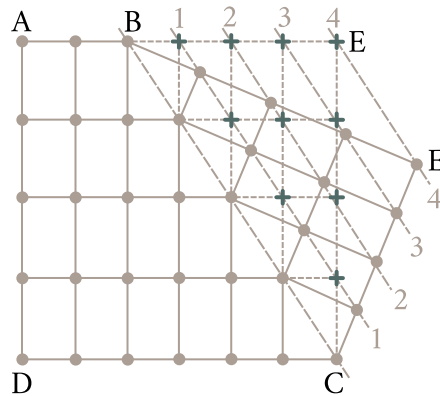


Figure 2.6: Twinning in a crystal: sign “+” denotes initial atomic positions in the twinning region.

absolute scale). In contrast to relaxation, which produces no visible changes in the crystal structure, recrystallization involves nucleation and growth of new crystals free from internal stresses. The nucleation of such crystals takes place in the first instance in the most deformed parts of the lattice, where much of the excess free energy is concentrated. In this way, a complete change in the microscopic structure of the crystal takes place with the crystal generally going over from the single state to the polycrystalline one. In the process of recrystallization the latent heat accumulated in the deformed crystal is given off in the form of heat.

§ 15. Mechanical twinning

Plastic deformation may also take the form of *twinning*, which is a process of step-by-step relative displacement of atomic planes parallel to the twinning plane by a fixed distance equal to a fraction of the lattice parameter. Figure 2.6 shows the diagram of twinning of the crystal AECDA. The area ABCDA is the undeformed part of the crystal, BECB is the part where twinning has taken place, and BC is the *twinning axis*. The positions of atoms before twinning are denoted by crosses. The plane passing through the twinning axis and separating the region of twinning from the undeformed part of the crystal is termed *twinning plane*.

Figure 2.6 shows that twinning results in the displacement of the atoms of the plane 11 relative to the twinning plane BC by a fraction of interatomic distance in the twinning direction. The plane 22 is displaced relative to the plane 11 by the same fraction of interatomic distance, the displacement relative to the twinning plane being twice as great. In other words, every atomic plane parallel to the twin-

ning plane is displaced in itself by a distance proportional to its distance from the twinning plane. As a result, the atoms in the twinned region assume positions that are mirror reflections of the positions in the undeformed part of the crystal in the twinning plane.

Twinning, in the same way as slipping, may take place only along specific crystallographic planes. In case of a face-centered cubic crystal this is the (112) plane, in case of a hexagonal close-packed crystal this is the (1012) plane, etc. For twinning to take place the tangential stresses should exceed some critical value. The process is a very rapid one and is usually accompanied by a characteristic crackle.

Because only negligible relative displacements of the neighbouring atomic planes are involved in the process of twinning it cannot result in a great residual deformation. For instance, a complete transition of a zinc crystal to the twinned form brings about only a 7.39% elongation. For this reason in crystals capable of plastic flow by means of slipping, twinning is responsible only for a negligible fraction of the total plastic deformation. In contrast to that, negligible deformation that precedes destruction of the valence crystals, in which slipping cannot take place, is due to twinning. In hexagonal crystals unfavourably oriented in relation to the external force twinning and subsequent reorientation of the crystal may result in appreciable residual deformations produced by the normal slipping process.

§ 16. Theoretical and real shear strengths of crystals

Shear is the principle mechanism of plastic flow in crystals. For a long time it was presumed that such shear takes place by means of a rigid displacement of one part of the crystal in relation to another simultaneously along the entire slip plane SS (Figure 2.7).

Let us make a rough estimate of the tangential stress needed to produce such shear.

The atoms of two adjacent parallel planes in an undeformed lattice occupy equilibrium sites corresponding to the minimum of the potential energy [Figure 2.7(a)]. The forces acting between them are zero. As one atomic plane is displaced relative to the other tangential stresses τ appear. They resist the shear and tend to bring back the original equilibrium state [Figure 2.7(b)]. If we assume the dependence of those stresses on the displacement to be sinusoidal (Figure 2.8), we shall be able to express the resistance to shear in the form

$$\tau = A \sin \left(\frac{2\pi x}{b} \right) \quad (2.11)$$

where x is the displacement of the atoms from their equilibrium positions, $b = a$

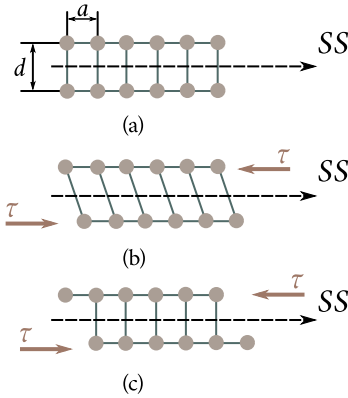


Figure 2.7: Diagram of rigid shear: (a) — equilibrium position of atoms in atomic planes adjoining the slip plane; (b) — gradual shift of one plane in relation to another caused by external stress τ ; (c) — lower atomic plane as a whole displaced by one interatomic distance in relation to the upper plane.

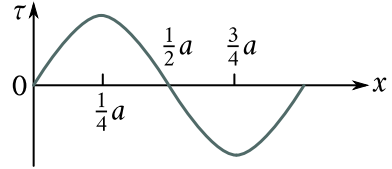


Figure 2.8: Variation of resistance to shear in the process of displacement of one part of the lattice in relation to another.

the interatomic distance in the slip plane, and A is a constant. For small displacements $\sin(2\pi x/b) \approx 2\pi x/b$ and therefore

$$\tau = A \left(\frac{2\pi x}{b} \right). \quad (2.12)$$

On the other hand, for small displacements Hooke's law is valid:

$$\tau = G \frac{x}{d} \quad (2.13)$$

where G is the shear modulus, and d the distance between the planes. From (2.12) and (2.13) we obtain $A = (b/d)(G/2\pi)$. Therefore

$$\tau = \frac{b}{d} \frac{G}{2\pi} \sin \left(\frac{2\pi x}{b} \right). \quad (2.14)$$

The maximum Value τ_{cr} of the tangential stress τ is attained for $x = b/4$ and this is assumed to be the theoretical strength:

$$\tau_{cr} = \frac{b}{d} \frac{G}{2\pi}. \quad (2.15)$$

Setting $b = d$, we obtain

$$\tau_{cr} = \frac{G}{2\pi}. \quad (2.16)$$

Hence, the critical shearing stress should be equal to about one tenth of the shear modulus. A more rigorous consideration of the nature of the bonding forces

acting between the atoms leads to a negligible correction factor. The minimum value that was obtained for τ_{cr} is $G/30$. Table 2.4 shows experimental and theoretical values of τ_{cr} for several metals.

A comparison of these figures shows that the real shear strength of crystals is some 3-4 orders of magnitude less than the theoretical value. This points to the fact that shear in crystals does not take place by means of a rigid relative displacement of atomic planes but by means of a mechanism involving the displacement of a comparatively small number of atoms at a time. The understanding of this fact led to the evolution of the dislocation theory of plastic flow of crystals.

§ 17. The dislocation concept. Principal types of dislocations

The dislocation theory of plastic flow assumes that the slipping process starts always at imperfections in the crystal structure and develops along the shear plane by means of a gradual motion of this imperfection which at a time involves only a limited number of atoms. Such imperfections are termed *dislocations*.

Edge dislocations. Suppose gliding took place in the crystal K in the plane ABCD in the direction of the vector \mathbf{b} involving the area AHED (Figure 2.9). The atomic planes on both sides from the slip plane AHED are displaced in relation to one another by the distance b in the slip direction. The boundary HE separating the area AHED, where slipping took place, from the area HBCE, where slipping has not yet taken place, constitutes an edge dislocation and the vector \mathbf{b} is termed the *Burgers vector*. It describes how far slipping has proceeded in the area AHED.

Table 2.4

Metal	τ_{cr} (10^7 Pa), experiment	G (10^7 Pa)	τ_{cr} (10^7 Pa), theory	
			$G/(2\pi)$	$G/30$
Cadmium	0.06	2640	420	88
Copper	0.10	4620	735	154
Iron	2.90	6900	1100	230
Magnesium	0.08	1770	280	59
Nickel	0.58	7800	1240	260
Silver	0.06	2910	459	97
Zinc	0.09	3780	600	126

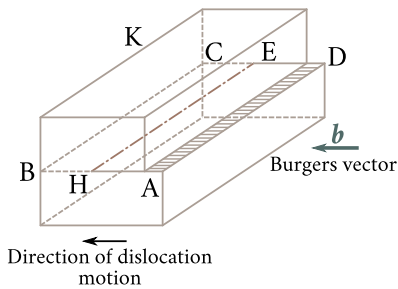


Figure 2.9: Shear that creates an edge dislocation. Shear took place only in region AHED of slip plane ABCD. Boundary HE is the edge dislocation.

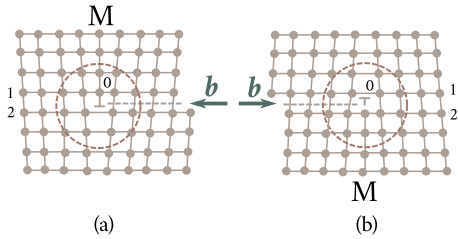


Figure 2.10: Arrangement of atoms in the plane perpendicular to dislocation line HE (see Figure 2.9). Dislocation occupies the region in which lattice atoms are displaced from their equilibrium sites (bounded by a circle): 0 — dislocation centre; (a) — positive dislocation; (b) negative dislocation.

Figure 2.10 shows the arrangement of atoms in a plane perpendicular to the dislocation line. As a result of the shift which took place over the area AHED the upper part of the lattice contains one atomic plane (plane OM) more than the lower. Because of that the atomic row 1 lying above the shear plane contains one atom more than the row 2 below this plane. The interatomic distances in the upper row near the point 0 (the dislocation centre) will accordingly be shorter than the normal value (the lattice is contracted), while the interatomic distances in the lower row near the point 0 will be longer (the lattice is extended). As the distance to the left or to the right, and up or down, from the dislocation centre 0 increases, the deformation of the lattice gradually subsides and at an appropriate distance from 0 in the crystal normal disposition of atoms is restored. However, in the direction perpendicular to the plane of the diagram the dislocation passes through the entire crystal or through a considerable part of it.

Thus, a feature of the edge dislocation is the existence of an “excess” atomic plane OM in some part of the crystal. Therefore the process of formation of such a dislocation may be imagined as that of pulling the lattice apart and inserting an additional atomic plane in it. Such plane is termed *extra plane*. If the plane is inserted into the upper part of the lattice, the edge dislocation is assumed to be positive [Figure 2.10(a)]. But if the extra plane is inserted into the lower part of the lattice, the dislocation is assumed to be negative [Figure 2.10(b)]. A dislocation whose Burgers vector is equal to the lattice parameter is called the *unit dislocation*. When a unit dislocation passes through a cross section of the crystal, one part of it shifts in relation to the other by a distance b . The motion of a positive dislocation to the left

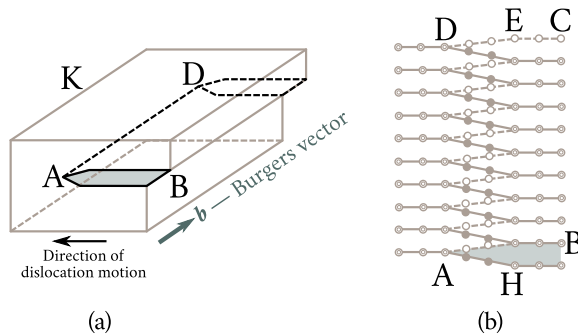


Figure 2.11: Formation of a screw dislocation: (a) — shear which produces the screw dislocation. It took place in the ABCD plane. Boundary AD is a screw dislocation; (b) — arrangement of atoms around the screw dislocation. Plane of drawing is parallel to slip plane. White circles denote atoms of the plane lying immediately above the slip plane, black circles denote atoms of the plane lying under the slip plane.

causes the same shift of parts of the lattice as a motion of a negative dislocation to the right [Figure 2.10(a,b)].

Screw dislocations. Suppose an incomplete unit shift is made in the crystal K in the direction of the vector \mathbf{b} over the area ABCD, as shown in Figure 2.11(a); AD is the boundary of the area that experienced the shift. In Figure 2.11(b) the white circles denote the atoms of the plane immediately above the slip plane and black circles the atoms of the plane below the slip plane. In the undeformed part of the crystal to the left of AD the atoms of those planes are arranged one on top of the other; therefore the black circles coincide with the white (this is shown by white circles with circles in the centre). In the right-hand part of the crystal, where the shift has covered one interatomic distance, that is, to the right of EH, the atoms of the planes discussed above are also arranged one on top of the other. However, in the narrow strip ADEH the atoms of the upper plane are displaced in relation to those of the lower plane the more the farther away they are from the boundary AD. This displacement results in a local deformation of the lattice, which became known as the *screw dislocation*; the boundary AD is termed *dislocation axis*. The origin of the term screw dislocation may be easily understood from Figure 2.12: the motion of the atom “a” towards the atoms “b, c, d, e”, etc. [Figure 2.12(a)] lying in the plane of the screw dislocation proceeds, as may be seen from Figure 2.12(b), along a spiral. A distinction is made between right and left screw dislocations (Figure 2.13); the motion of both in opposite directions results in a shift in one direction.

Comparing Figures 2.9 and 2.11(a), we see that in contrast to the edge dislocation, which is perpendicular to the Burgers vector \mathbf{b} , the screw dislocation is

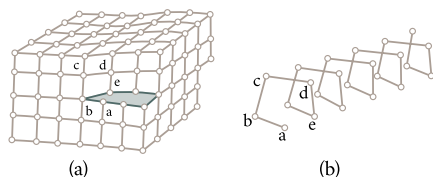


Figure 2.12: Explaining the origin of the “screw dislocation”: (a) — arrangement of atoms in a screw dislocation; (b) — atom “a” moves towards atoms “b, c, d, e”, etc. constituting the screw dislocation along a spiral.

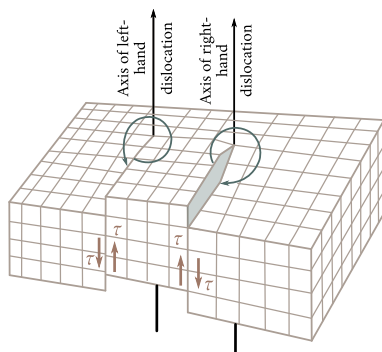


Figure 2.13: Arrangement of atoms in the plane perpendicular to dislocation line HE (see Figure 2.9). Dislocation occupies the region in which lattice atoms are displaced from their equilibrium sites (bounded by a circle): 0 — dislocation centre; (a) — positive dislocation; (b) negative dislocation.

parallel to it. The motion of the edge dislocation is in the direction of the Burgers vector \mathbf{b} , and the motion of the screw dislocation is in the direction perpendicular to it.

Recently, experimental methods for direct observation of dislocations have been developed. Figure 2.14(a) shows a schematic diagram of an electron micrograph of a thin film of platinum phthalocyanine and Figure 2.14(b,c) an electron micrograph of a copper sulphide crystal obtained with the aid of a special procedure. Dark stripes on the micrographs are the traces of the atomic planes, which in platinum phthalocyanine are arranged at a distance of 12 \AA from one another and in copper sulphide at a distance of 1.88 \AA . The micrographs distinctly show the extra planes which terminate inside the crystal and form edge dislocations.

Figure 2.14(d) shows an optical micrograph of a decorated screw dislocation in a CaF_2 crystal. The method of decoration as used for transparent crystals consists in the precipitation along their dislocation cores of impurity atoms, which make the dislocation visible in an optical microscope. The striking agreement between those pictures and the theoretical concepts as set out in Figures 2.10 and 2.12 is worthy of admiration. Points of exit of dislocations on the crystal surface may be detected with the aid of etching. When a crystal is etched in a specially selected etch, the parts of the crystal where the lattice is most deformed dissolve more readily because the atoms in those parts possess an excess energy and are chemically more

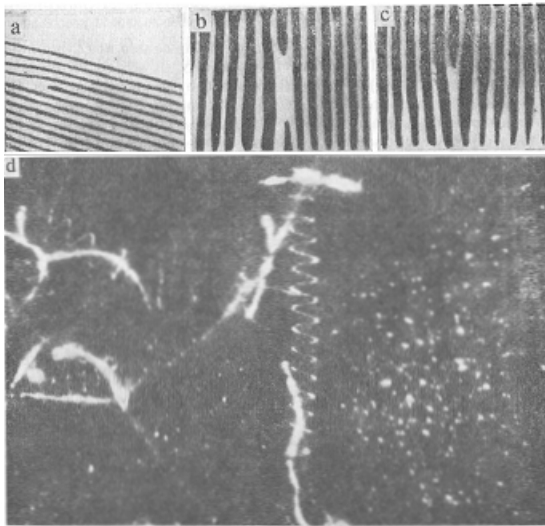


Figure 2.14: Observation of dislocations in an electron microscope: (a) — schematic diagram of an electron micrograph of a thin platinum phthalocyanine film (dark lines are atomic traces); (b), (c) — electron micrograph of a copper sulfide crystal (dark lines are traces of atomic planes); (d) — screw dislocation in a CaF_2 crystal obtained by decoration method.

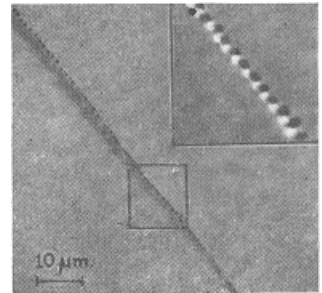


Figure 2.15: Etch pits on germanium surface. Dark points along the grain boundary are points of exit of dislocations.

active. The places of exit of dislocations on the crystal surface are just such parts. Figure 2.15 shows a photograph of an etched germanium surface. The dark patches are the points of exit of dislocations.

§ 18. Forces needed to move dislocations

Suppose there is a positive dislocation with the centre at 0, bounded at points “a” and “k” and lying in the plane S in which slipping is possible [Figure 2.16(a)]. In equilibrium the force with which the lattice acts on the dislocation is zero. This may easily be seen from the roller model shown in Figure 2.17. The structure of the upper row of rollers which normally occupy recesses between the rollers of the lower row was deformed so that the section AB which previously contained 6 rollers now contains only 5. Such deformation gives rise to forces which tend to return the rollers 1, 2, 4, 5 to their stable equilibrium positions (the forces F_1, F_2, F_3, F_4, F_5). The forces applied to rollers 7, 5 and 2, 4 are equal in magnitude and opposite in direction. Therefore, if the rollers of the upper row are interconnected by means

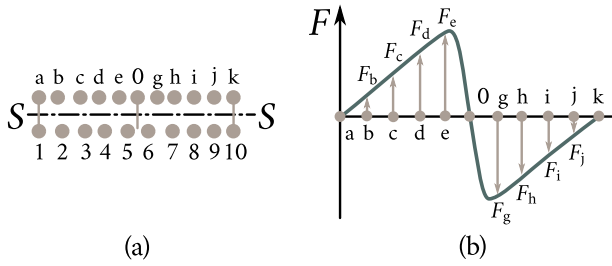


Figure 2.16: Calculating the force needed to move a dislocation: (a) — region of positive dislocation in crystal; 0 is dislocation centre, “a” and “k” are dislocation boundaries, S is the slip plane; (b) — forces needed to move an atom in the dislocation region (forces applied to atoms equidistant from the dislocation centre are equal in magnitude and opposite in direction).

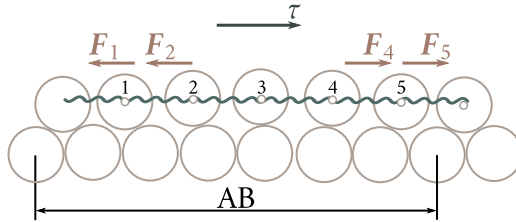


Figure 2.17: Roller model of an edge dislocation. Forces applied to “atoms” 1, 5 and 2, 4 are equal in magnitude and opposite in direction.

of an elastic spring acting as a bond between them, the forces F_1 and F_5 , F_2 , F_4 will be mutually compensated and the system will be in a state of equilibrium.

The same situation occurs in the case of a dislocation schematically shown in Figure 2.16(b); the forces acting on atoms of the upper row occupying positions symmetrical with respect of the dislocation centre 0 are equal in magnitude but opposite in direction (the forces $F_b = F_j$, $F_c = F_i$, $F_d = F_h$, $F_e = F_g$). Therefore the resultant force is zero and the dislocation is in a state of equilibrium. If, however, the dislocation moves a little in the slip plane the symmetrical arrangement of the atoms in respect of the dislocation centre will be disturbed giving rise to a force which resists the motion of the dislocation. It is evident from Figure 2.17 that this force cannot be great since the movement of the rollers 1 and 2 to their new equilibrium position is to a large extent the result of the action of the forces exercised on them by the rollers 4 and 5, which also strive to occupy positions of stable equilibrium. Calculations show the tangential stress needed to move the dislocation to be equal to

$$\tau_0 = \frac{2G}{(1-\nu)} \exp \left[-\frac{2\pi b}{d(1-\nu)} \right] \quad (2.17)$$

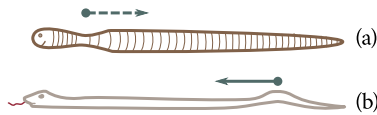


Figure 2.18: Dislocation mechanism of motion of (a) an earth-worm and (b) a snake.

where G is the shear modulus, ν the Poisson ratio, b the interatomic distance, and d the distance between adjacent slip planes. The stress τ_0 is the theoretical value of the critical shearing stress. Setting $b = d$ and $\nu = 0.3$, we obtain $\tau_0 = 3 \times 10^{-4} G$. Within an order of magnitude this coincides with the experimental values of τ_{cr} . Thus, the theory of dislocations resolves the contradiction between the theoretical and the experimental values of shear strength of crystals.

The mechanism of motion by means of dislocations is quite frequent in nature. Snakes, worms, and shellfish move, because they generate dislocations. The motion of an earth-worm begins with the formation of an “extension” dislocation near the neck. The dislocation subsequently spreads along the body to the tail [Figure 2.18(a)]. In contrast, the motion of most snakes involves the formation of a “contraction” dislocation near the tail and its motion towards the head [Figure 2.18(b)].

§ 19. Sources of dislocations. Strengthening of crystals

The dislocations in a real crystal are formed in the process of its growth from the melt or from a solution. Figure 2.19(a) shows the boundaries of two blocks growing towards each other. The blocks make a small angle φ between themselves. As the blocks fuse together, some of the atomic planes do not spread through the entire crystal but terminate at block boundaries. Those are the places where dislocations are formed [Figure 2.19(b)]. The same situation occurs in the process of fusion of differently oriented grains in a polycrystalline sample. Since the block and grain boundaries in real solids are very extensive, the number of dislocations in them is enormous—as many as 10^{12} dislocations per square metre can be counted in well annealed metals. After cold working (rolling, drawing, etc.) dislocation densities rise to 10^{15} m^{-2} to 10^{16} m^{-2} . Those dislocations accumulate almost the entire energy absorbed by the metal in the process of plastic deformation.

Vacancy clusters may also serve as sources of dislocations in an undeformed crystal. Figure 2.20 shows an example of the formation of a positive and a negative dislocation from a cluster of vacancies.

The shear process in a crystal in response to an applied external force is, in effect, the motion of dislocations in the slip planes and their emergence through the crystal surface. Should only the dislocations already present in the crystal be

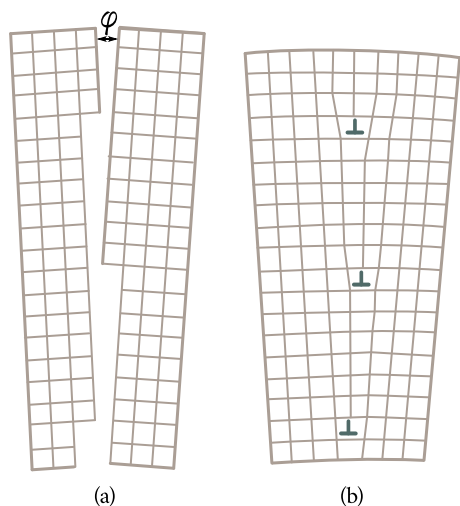


Figure 2.19: Formation of dislocations at block boundaries: (a) — two blocks growing towards each other at an angle φ ; (b) — dislocations appearing when the blocks fuse together.

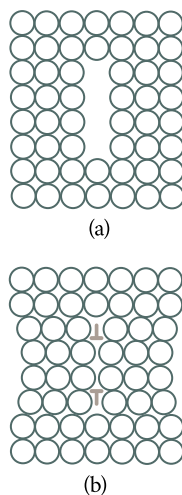


Figure 2.20: Dislocations formed from vacancy clusters: (a) — vacancy cluster in crystal; (b) — positive and negative dislocations formed from this cluster.

responsible for this process, plastic deformation would lead to their exhaustion and to the perfection of the crystal. This is in contradiction with experience, which demonstrates that as deformation grows the lattice does not become more perfect. In fact, just the opposite is true: the density of dislocations grows in the process. It is an established fact that dislocations responsible for plastic deformation are generated in the shear process itself by the action of the external force applied to the crystal. One such generation mechanism was discovered in 1950 by F. C. Frank and W. T. Read. For the purpose of better understanding this mechanism let us consider soap bubble formation with the aid of a tube (Figure 2.21). After the end of the tube has been immersed in a soap solution a flat film remains that closes the tube's orifice. As the air pressure in the tube is increased, the film swells and passes through the stages 1, 2, 3, 4, etc. Until it assumes the shape of a hemisphere (stage 3) its state is unstable: as pressure falls the film contracts striving to return to the original state 1. After the bubble has passed stage 3 the state of the bubble changes: it is now capable of evolution not only at a constant but also at a gradually decreasing pressure until it leaves the end of the tube. After the first bubble the second begins to be formed, followed by the third, etc.

Now let us discuss the operation of the Frank-Read source. Figure 2.22(a) shows an edge dislocation DD' in a slip plane; points D and D' are fixed and do not take part in the motion of the dislocation. Dislocations may be anchored at the points of

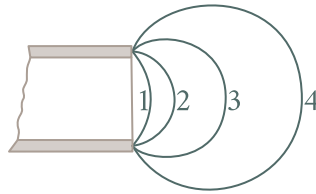


Figure 2.21: Process of formation of a soap bubble.

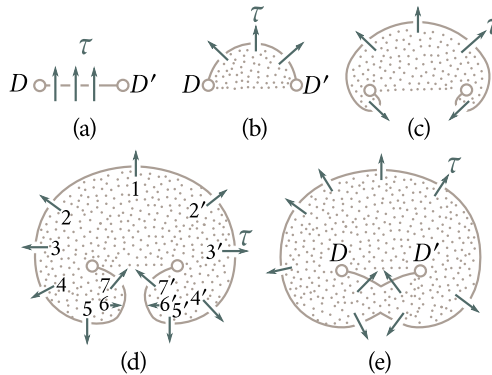


Figure 2.22: Operating sequence of a Frank-Read source: (a) — initial position of dislocation DD' , (b) — acted upon by external force the dislocation bends and assumes the shape of a semicircle; (c), (d) — further symmetric development of the dislocation loop; (e) — formation of external closed dislocation loop spreading across the crystal and of internal dislocation DD' returning to the original position.

intersection with other dislocations, at impurity atoms, etc. Under the action of an external stress τ the dislocation starts bending in the same way as was the case with the soap film and at some time assumes the shape of a semicircle [Figure 2.22(b)]. Just like the soap film the dislocation can continue to bend only if τ grows steadily until it assumes the shape of a semicircle. Its subsequent evolution takes place by itself and results in the formation of two loops [Figure 2.22(c)], which after meeting at point C [Figure 2.22(d)] divide the dislocation in two: an external one, which closes forming an external circle [Figure 2.22(e)], and an internal one, which returns to the original position DD' . The external dislocation grows until it reaches the surface of the crystal and results in an elementary shift; the internal dislocation having returned to the initial position DD' begins again to bend under the action of the applied force and to grow in the manner described above. Such process may be repeated any number of times eventually leading to a noticeable shift of one part of the crystal in relation to another in a particular slip plane.

Low shear strength of crystals is due to the presence of innate dislocations and

to the generation of others in the process of Shear. On the other hand, it is an established fact that the crystal is strengthened in the process of plastic deformation accompanied by the growth in the number of defects. The essence of such strengthening is the interaction of dislocations with each other and with other types of lattice defects causing their motion in the lattice to be obstructed.

Interaction of dislocations. Every dislocation, being the cause of elastic stresses of the lattice, creates a force field around itself which may be described by the values of the tangential τ and normal σ stresses at every point. When another dislocation enters this field, forces begin to act which strive to bring the dislocations together or to move them apart. Dislocations of like signs lying in one plane are repelled, while those of opposite signs are attracted. This is the reason why, as dislocations are accumulated in a definite plane, the crystal's resistance to shear is increased and the crystal is strengthened.

Surmounting of obstacles. Suppose a dislocation when moving in a slip plane under the action of tangential stresses τ runs into a stationary obstacle D, for instance, an intersection with some other dislocation, an impurity atom, or some other type of defect. Figure 2.23 shows the method by means of which dislocation AB could, theoretically, surmount obstacle D: as the dislocation approaches D (positions 1, 2, 3) it gradually bends forming a loop that envelops the obstacle. Behind the obstacle the loop closes and the dislocation A'B' again becomes a straight line. Figure 2.24 shows a photograph illustrating a case when a dislocation runs into a stationary obstacle (dark lines represent dislocations decorated by etching). The similarity in the pictures leaves not a trace of doubt as to the validity of the theoretical pattern shown in Figure 2.23.

In passing around the obstacle, the length of the dislocation and the deformation of the lattice are increased, which requires additional work to be performed. Therefore, the resistance to the motion of the dislocation in the interval where it has to surmount a defect is much greater than in other parts of the lattice. This is the essence of the fact that defects strengthen a crystal. The growth in the number of dislocations in the crystal with greater plastic deformation increases the number of obstacles at points of their intersection, which is the cause of strengthening brought about by plastic deformation. Impurity atoms have a similar effect: they create local lattice imperfections and thereby hinder the motion of the dislocations, with the result that the crystal's resistance to shear is increased. Block and grain boundaries and foreign inclusions in the lattice are especially effective in hindering the motion of the dislocations. They sharply increase the resistance to the motion of dislocations and greater stresses are required to overcome their effect. The phenomenon of strengthening in the process of cold working, in the process of introducing impurity atoms (doping), and in the process of formation

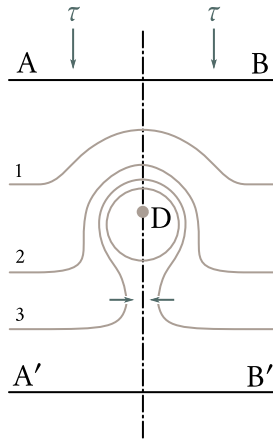


Figure 2.23: Schematic representation of an edge dislocation surmounting an obstacle: AB — shape of edge dislocation away from the obstacle D; 1, 2, 3 — gradual bending of the dislocation as it approaches D and closing of the newly formed loop behind the obstacle; A'B' — straightening of the dislocation far away from the obstacle.



Figure 2.24: Microphotographs of a chromium grain. Dark lines are etched dislocations ($\times 2000$).

of inclusions (tempering, aging, etc.) is widely used in practice to improve mechanical properties of engineering materials. This method enabled the strength of the materials to be increased from 6 to 8 times in the last 40 years.

§ 20. Brittle strength of solids

The destruction of solids may be of one of two principal types: of the *brittle* and of the *plastic*, or *viscous*, types.

Brittle destruction takes place if the tensile strength of the material is below the elastic limit. Such a material experiences only elastic deformation prior to destruction. No irreversible changes take place in such a material before it breaks down.

In the ductile materials the elastic limit is not only below the tensile strength but also below the yield stress. Because of that the destruction process is preceded by an appreciable plastic deformation, which prepares the subsequent destruction process. In this case strength, being a typical kinetic parameter, is strongly dependent on the time the destructive stress is applied.

To begin with, let us discuss brittle strength of solids.

Theoretical strength of solids. There have been numerous attempts to calculate the strength of solids on the basis of molecular interaction in them. The

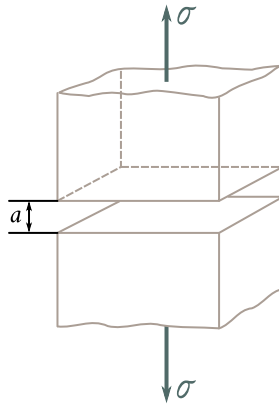


Figure 2.25: Calculating theoretical strength of solids after Polanyi (explanation in text).

strength σ_0 thus calculated is termed *theoretical strength*.

Here is a glance at some of the methods of estimating σ_0 .

Polanyi's method. The simplest method of estimating the strength of solids theoretically is due to M. Polanyi. Its essence is as follows.

Suppose a tensile stress σ is applied to a rod of a cross-sectional area of 1 m^2 (Figure 2.25). This stress increases the distance between the atomic planes. It is assumed that for destruction to take place a stress σ_0 able to increase the distance between the atomic planes by a value of the order of the lattice parameter a should be applied. The work needed to move an atomic plane a distance a away from the neighbouring plane is assumed to be equal to $\sigma_0 a$. It is further assumed that this work is transformed into the free energy of two new surfaces with a total area of 2 m^2 formed as a result of the breakup, the free energy being equal to 2α , where α is the surface energy ("surface tension") of the solid. Hence,

$$\sigma_0 a = 2\alpha$$

and the theoretical strength is

$$\sigma_0 = \frac{2\alpha}{a}. \quad (2.18)$$

For copper $\alpha \approx 1.7 \text{ J m}^{-2}$, $a = 3.6 \times 10^{-10} \text{ m}$, and $\sigma_0 \approx 10^{10} \text{ Pa}$, for silver $\alpha \approx 1.14 \text{ J m}^{-2}$, $a = 4 \times 10^{-10} \text{ m}$, and $\sigma_0 = 0.6 \times 10^{10} \text{ Pa}$.

Determination of σ_0 from the heat of sublimation. The energy equal to the heat of sublimation Q_s is required for the evaporation of a mole of a solid. For the evaporation of one molecular layer of the area of 1 m^2 the required energy W is a fraction of Q_s equal to the ratio of the mass of this layer m to the molar mass M :

$$W = \frac{Q_s}{m}.$$

But

$$m = N_s \mu, \quad M = N_A \mu$$

where μ is the molecular weight, $N_A = 6.023 \times 10^{23} \text{ mol}^{-1}$ the Avogadro number, and N_s the number of molecules per square metre of the solid's surface.

For an intermolecular distance of a the area per molecule is approximately equal to a^2 and the number of molecules per square metre $N_s \approx a^2$. Therefore,

$$W = Q_s \frac{N_s}{N_A} \approx \frac{Q_s}{N_A a^2}.$$

Should the assumption be made that the evaporating molecules loose contact with the solid's surface when they are a distance of the order of the lattice parameter a away from it, we would obtain for the force needed to tear away an entire surface layer as a whole

$$\sigma_0 \approx \frac{W}{a} = \frac{Q_s}{N_A} \frac{1}{a^2} \quad (2.19)$$

σ_0 is assumed to be the theoretical strength of the solid.

For copper $Q_s = 3 \times 10^5 \text{ J mol}^{-1}$, $a = 3.6 \times 10^{-10} \text{ m}$, and $\sigma_0 \approx 10^{10} \text{ Pa}$. Similar calculations lead to the following results: for iron $\sigma_0 \approx 2.3 \times 10^{10} \text{ Pa}$, for aluminium $\sigma_0 \approx 0.6 \times 10^{10} \text{ Pa}$, and for silver $\sigma_0 \approx 0.6 \times 10^{10} \text{ Pa}$.

Calculating σ_0 from the forces of molecular interaction. Finally, let us discuss the method of calculating theoretical strength of solids from the forces of molecular interaction. Figure 2.26 shows the dependence of the potential energy $U(x)$ and the force of interaction between the particles $f(x)$ on the distance x between them. Since it is not easy to determine the exact law governing $f(x)$, the practice is to approximate this dependence by various functions. For instance, M. Polanyi and E. Orowan used the approximation in the form of a half of a sinusoid:

$$f(x) = f_{\max} \sin \left(\frac{2\pi x}{c} \right). \quad (2.20)$$

When a body of cross-sectional area of 1 m^2 is slowly torn in two, the required force is $\sigma = f N_s$, where N_s is the number of particles per square metre of the cross section. Substituting f from (2.20) we obtain

$$\sigma = \sigma_0 \sin \left(\frac{2\pi x}{c} \right) \quad (2.21)$$

where $\sigma_0 = f_{\max} N_s$ is the theoretical strength of the body.

For small displacements relation (2.21) may be rewritten in the form

$$\sigma = \frac{\sigma_0 2\pi x}{c}.$$

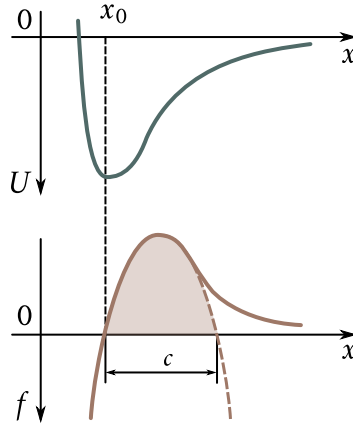


Figure 2.26: Calculating theoretical strength from forces of molecular interaction (explanation in text).

On the other hand, for small displacements Hooke's law is valid:

$$\sigma = \frac{Ex}{c}.$$

Equating the right-hand sides of these equations, we obtain

$$\sigma_0 \approx \frac{E}{2\pi} \approx 0.1E. \quad (2.22)$$

Calculations show that a more accurate estimate of the nature of the bonding in solids results only in a negligible correction to (2.22).

Comparing the values of theoretical strength σ_0 calculated with the aid of various methods we see that all of them yield nearly the same result whose order of magnitude is $0.1E$. Therefore it may be legitimately assumed that

$$\sigma_0 \approx 0.1E. \quad (2.23)$$

This is an enormous figure of the order of 10^9 Pa to 10^{10} Pa.

Real (technical) strength of solids. The strength of real crystals and solids used for technical purposes is termed *real*, or *technical*, strength σ_r . Table 2.5 shows the values of the elasticity modulus E , of theoretical strength $\sigma_0 \approx 0.1E$, of technical strength σ_r , and of the ratio σ_0/σ_r for some industrial materials.

It follows from the data of Table 2.5 that the technical strength of solids is from 2 to 3 orders of magnitude less than their theoretical strength.

At present there is a general agreement that such discrepancy between σ_0 and σ_r is due to the presence of defects in real solids of various types, in particular of microscopic cracks which reduce the strength of solids. This is accounted for by the so-called *Griffith theory*. Let us calculate the technical strength using this

theory.

We take a sample in the form of a thin plate and apply a tensile stress σ to it [Figure 2.27(a)]. The density of elastic energy in such an elastically extended sample is $\sigma^2/(2E)$ ¹.

Now let us imagine that a transverse microscopic crack of the length l running through the entire thickness δ of the sample has developed in it. The appearance of the crack is accompanied by the formation of a free surface $S \approx 2l\delta$ inside the sample and by an increase in the sample's energy by the amount $\Delta U_1 \approx 2l\delta\alpha$ (α is the free surface energy of the sample per unit area). On the other hand, the formation of a crack relieves the elastic stress from the volume $V \approx l^2\delta$ of the sample, whereby its elastic energy is reduced by the amount $\Delta U_2 \approx l^2\delta\sigma^2/(2E)$. The total change in the energy of the sample $W(l)$ brought about by the appearance of a crack in it is

$$W(l) = 2l\delta\alpha - l^2\delta\frac{\sigma^2}{2E}. \quad (2.24)$$

Figure 2.27(b) shows the dependence of W on the length l of the crack. It has a maximum where its derivative vanishes: $dW/dl = 2\delta\alpha - l\delta\sigma^2/E$. Denote the length of the crack corresponding to the maximum energy by l_{cr} . We obtain from the last relation

$$l_{cr} = \frac{2\alpha E}{\sigma^2}. \quad (2.25)$$

¹Indeed, the relative deformation in a sample under stress σ is $\varepsilon = \sigma/E$, the absolute deformation $\Delta L = \varepsilon L$ (L being the length of the sample). The work performed by the stress σ to extend the sample by ΔL is $(1/2)\sigma S \Delta L = \sigma^2 SL/(2E) = \sigma^2 V/(2E)$ (S is the cross-sectional area and V the volume of the sample). This work is transformed into the elastic energy of the sample of volume V . Therefore, the specific volume density of the elastic energy is $\sigma^2/(2E)$.

Table 2.5

Substance	Elasticity modulus E (10^7 Pa)	Theoretical strength $\sigma_0 \approx 0.1E$ (10^7 Pa)	Technical strength σ_r (10^7 Pa)	σ_0/σ_r
Aluminium	6000	600	9.0	65
Copper	12000	1200	23	50
Glass	8000	800	8.0	100
Iron	21000	2100	30	70
Rock salt	4000	400	0.5	800
Silver	8000	800	18	45

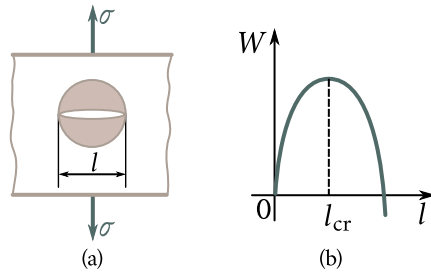


Figure 2.27: The Griffith theory of calculating the real strength of solids (explanation in text).

It may be seen from Figure 2.27(b) that as long as the length l of the crack remains below the critical value l_{cr} , energy is needed for it to develop. On the other hand, starting with $l = l_{cr}$ further extension of the sample results in a reduction in its energy. Therefore, it takes place spontaneously with the brittle destruction of the sample as the final result.

Hence, the technical strength of solids having microscopic cracks should be calculated according to the Griffith theory from relation (2.25):

$$\sigma_r \approx \left(\frac{2\alpha E}{l} \right)^{1/2} \approx \beta \left(\frac{\alpha E}{l} \right)^{1/2}. \quad (2.26)$$

This result was subsequently verified by many investigators for various quite different methods of applying loads to the sample. A negligible difference was observed only in case of the numerical coefficient β .

Should we substitute the values of α, E, σ_r for copper ($\alpha \approx 1.7 \text{ J m}^{-2}$, $E = 1.2 \times 10^{11} \text{ Pa}$, $\sigma_r = 1.8 \times 10^8 \text{ Pa}$) into (2.26), we would obtain $l \approx 8 \times 10^{-6} \text{ m}$. Approximately the same values of l may be obtained for other solids.

It follows that for the strength of the solids to be reduced from the theoretical value to the value of the technical strength microscopic cracks of the order of several micrometers in length should develop in them up to the moment preceding their destruction. Many factors may be the cause of such cracks.

The cracks may be produced in the course of the production of the solid, especially in the course of its mechanical processing. A proof of this is, in particular, a significant dependence of the strength of the sample on its dimensions, especially in the small dimensions range. Thus the strength of a glass filament of $2.5 \mu\text{m}$ in diameter is almost 100 times that of a massive sample. The explanation is that as the dimensions of the sample are reduced so too is the probability of a large crack responsible for low strength appearing in it. Such dependence of the strength on the dimensions of the sample became known as the *scale factor*. The cracks may be

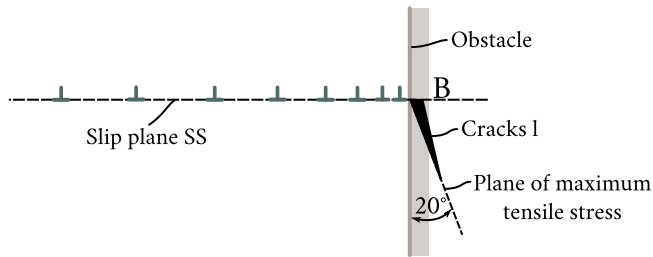


Figure 2.28: Formation of a crack near a dislocation pile-up.

the result of a large number of vacancies merging together.

Figure 2.28 shows a dislocation mechanism of crack production. Dislocations of a similar sign move in slip plane SS and meeting obstacle B begin to accumulate in its vicinity. Large stresses able to produce cracks l may develop at the head of this dislocation pile-up.

§ 21. Time dependence of the strength of solids

The theory of strength based on the condition (2.26) and discussed above describes actually the final stage of the destructive process when the body already contains cracks able to cause brittle rupture.

However, the initial stages of the destructive process during which the cracks originate and grow to attain critical dimensions l_{cr} are also important. This process is a gradual one and takes time τ to be completed. The time τ that it takes for the destructive process to develop from the moment the load is applied to the body to the moment of rupture is termed the *durability* of the material.

The firsts experiments aimed at investigating durability were carried out by S. N. Zhurkov and G. M. Bartenev with coworkers. They also developed modern notions about the physical nature of durability.

It was established by experiments that the durability τ , the tensile stress σ , and the absolute temperature T are related by the expression:

$$\tau = \tau_0 e^{(U_0 - \gamma\sigma)/(k_B T)} \quad (2.27)$$

where τ_0 , U_0 , and γ are constants dependent on the nature and structure of the body.

For $T = \text{constant}$, formula (2.27) may be rewritten in the form

$$\tau = A e^{-\beta\sigma} \quad (2.28)$$

where $A = \tau_0 e^{U_0/(k_B T)}$, and $\beta = \gamma/(k_B T)$.

Formulae (2.27) and (2.28) were tested on a great number of different materials

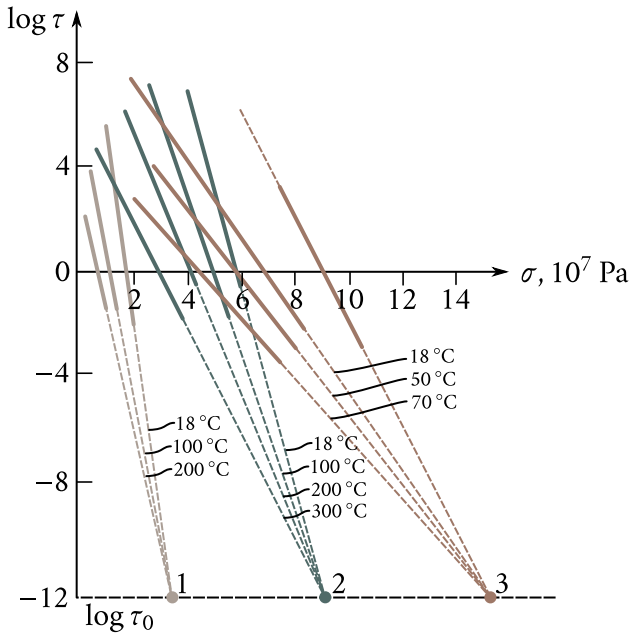


Figure 2.29: Durability versus stress for aluminium (1), Plexiglas (2), and silver chloride (3).

(metals, polymers, haloid compounds, etc.) in an 8 to 10 order of magnitude range of the values of τ and in a wide range of the values of T .

Figure 2.29 shows the dependence of the durabilities τ of aluminium (1), Plexiglas (2), and silver chloride (3) on the applied stress σ at various temperatures expressed in the $\log \tau$ versus σ coordinate system. It may be seen from Figure 2.29 that the dependence $\tau(\sigma)$ in semilogarithmic coordinates is well represented by a straight line. A family of such straight lines obtained for a given material at different temperatures resembles a fan with the apex at some point called pole. It follows from Eq. (2.27) that τ will be independent of T and that the straight lines $\log \tau(\sigma)$ at different temperatures will intersect at one point (at the pole) only if $U_0 - \gamma\sigma = 0$; but in that case $\log \tau = \log \tau_0$. Hence, the pole should be at a distance $\log \tau_0$ below the σ -axis.

It is evident from Figure 2.29 that the poles for all the materials tested lie practically on the same straight line parallel to the σ -axis. This means that τ_0 is approximately the same for all the materials. Experiments show it to be of the order of 10^{-12} s to 10^{-13} s, that is, close to the period of oscillations of atoms in solids.

Let us take the logarithm of (2.27):

$$\log \tau = \log \tau_0 + \frac{(U_0 - \gamma\sigma)}{k_B T} = \log \tau_0 + \frac{U}{k_B T}, \quad U = U_0 - \gamma\sigma. \quad (2.29)$$

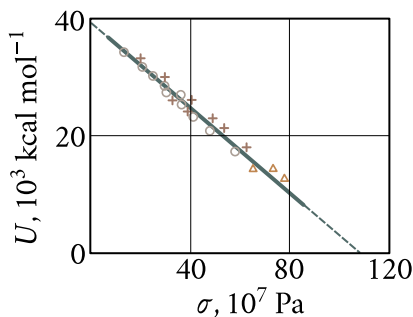


Figure 2.30: Activation energy of rupture of viscose fibre at different temperatures (triangles, $-76\text{ }^{\circ}\text{C}$; circles, $+20\text{ }^{\circ}\text{C}$; crosses, $+80\text{ }^{\circ}\text{C}$) versus stress.

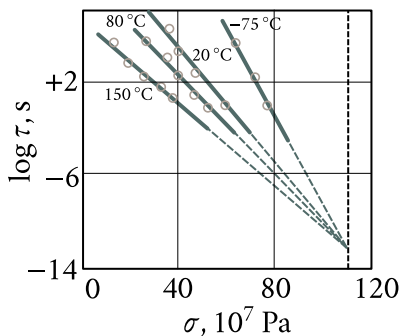


Figure 2.31: Durability of viscose fibre at different temperatures versus stress.

Measuring the dependence of $\log \tau$ on $1/T$ for constant values of σ , we can determine U for various values of the stress σ experimentally; the dimensions of U are that of energy and because of this it is termed *activation energy of the destructive process*. Figure 2.30 shows the dependence of the activation energy of rupture of viscose fibre on stress for various temperatures. It may be seen that U is independent of T and is determined solely by σ ; for $\sigma = 0$ the maximum value of U is $U_0 \approx 40 \text{ kJ/mol}$; for a stress $\sigma \approx 107 \times 10^7 \text{ Pa}$ we see that $U = 0$. Figure 2.31 shows that for $\sigma \approx 107 \times 10^7 \text{ Pa}$ a practically instantaneous rupture of viscous fibre takes place (during the time τ_0), no matter what its temperature is.

Meticulous experiments carried out by S. N. Zhurkov with coworkers and by other investigators on a variety of materials demonstrated that for metals U_0 is quite close to the sublimation energy Q_s and for polymers to the thermal destruction energy Q_d . Table 2.6 shows the values of U_0 , Q_s , and Q_d for some materials. It may be seen that U_0 coincides either with Q_s or with Q_d with a high degree of accuracy.

Universal validity of the dependence thus obtained merits the conclusion that the process of destruction of a solid is one of a kinetic nature (that is, develops in time) and its origin is the same for all solids. Modern notions of the physical mechanism of this process are set out below.

The atoms in a solid take part in thermal oscillations with a period of $\tau_0 \approx 10^{-12} \text{ s}$ to 10^{-13} s . Thermal fluctuations from time to time result in the rupture of chemical bonds. The probability of this process depends on the height of the potential barrier of destruction U and on the temperature T . This probability increases with the rise in T and the decrease in U . In the absence of external stress σ

the energy required to break a bond is equal to the energy of the bond itself. This is the reason why the height of the potential barrier U_0 obtained from experiments in mechanical destruction of solids turned out to be equal to the sublimation heat of metals and to the thermal destruction energy of polymers.

The stresses induced in a body reduce the height of the potential barrier from U_0 to $U_0 - \gamma\sigma$ and thus increase the probability of rupture of the bonds and, consequently, the number of ruptured bonds per unit volume.

The formation of submicroscopic volumes in which the bonds have been broken and their merging results eventually in the nucleation and development of cracks. When the length of the cracks attains a critical value, the body breaks up under the applied stress. The higher is the stress the lower the activation barrier $U_0 - \gamma\sigma$ and the greater the rate of bond rupture; therefore it takes less time for the destructive process to develop, that is the less should be the durability of the body. This is exactly what is observed in practice.

From the above point of view the destruction of solids should take place at any stresses provided the time they act is long enough. But in that case it is not easy to understand why bridges and other installations built many centuries ago and carrying loads all that time still remain intact.

To explain this fact we again turn to Figure 2.29. We see that the lower the temperature the weaker the load dependence of durability is. This dependence is practically nill at sufficiently low temperatures. For glasses and metals with a high melting point already room temperatures are low enough. Because of that their strength is actually a unique characteristic of the material. In all other conditions it is not justified to speak of strength without mentioning the time during which the material is to work under load. Thus, industrial products made of Plexiglas

Table 2.6

Substance	Activation energy of destruction, U_0 (10^5 J mol ⁻¹)	Sublimation energy Q_s (10^5 J mol ⁻¹)	Thermal destruction energy, Q_d (10^5 J mol ⁻¹)
Aluminium	2.16	2.2	
Nickel	3.48	3.4	
Nylon	1.8		1.72
Platinum	4.8	5.1	
Polymethyl methacrylate	2.16		2.1-2.2
Polyvinyl chloride	1.4		1.28
Silver	2.56	2.72	
Teflon	3.0		3.0-3.1
Zinc	1.0	1.08	

during a year's service can endure loads not exceeding 30% of their short-time strength; steam turbine blades working at high temperatures are calculated for strength always with account taken of their durability.

§ 22. Methods of increasing the strength of solids

The nucleation mechanism of breaks in continuity and the mechanism of crack growth are both greatly influenced by the atomic structure of solids. Therefore, strength is a structure-sensitive characteristic of such bodies.

Stresses in crystals occasion the production of dislocations and their motion in slip planes. In this way, plastic shifts resulting in plastic deformations are realized. Meeting impurities, grain and block boundaries, interceptions of slip planes, etc., the dislocations lose their mobility and the crystal is hardened. As was mentioned above, stresses may develop at the head of a dislocation pile-up capable of causing cracks.

To increase the strength of such bodies it is necessary to retard the production of dislocations and the nucleation and growth of cracks.

This can be done by two methods.

(i) By producing imperfection-free crystals free from internal stress sources, which in the long run cause the nucleation of cracks.

This method has up to the present been realized only in the filament type crystals known by the name of "whiskers". They are single crystals grown under special conditions using the method of decomposition or reduction of appropriate chemical compounds, the method of condensation of vapours of pure metals at an appropriate temperature in hydrogen or in an inert gas, and the method of electroplating metals from a solution onto electrodes of extremely small dimensions. The filament-type crystals are usually 2 mm to 10 mm long and $5\text{ }\mu\text{m}$ to $50\text{ }\mu\text{m}$ thick.

A striking property of such crystals is that their mechanical parameters are extremely high. Their strength turned out to be close to the theoretical strength of solids. Thus, the strength of iron whiskers is about 1.34×10^{10} Pa, of copper whiskers about 3×10^9 Pa, and of zinc whiskers 2.3×10^9 Pa, while the strength of normal samples made from those metals is 3×10^8 Pa, 2×10^8 Pa, and 1.8×10^8 Pa, respectively.

Filament-type crystals of iron experience only elastic deformation reaching an enormous figure of the order of 5%-6%, after which brittle destruction occurs. Note that in normal iron noticeable plastic flow starts already at a deformation of $\varepsilon \approx 0.01\%$.

The unusually high mechanical parameters of the filament crystals are due to their ideal internal structure. Such crystals contain practically no dislocations, are

exceptionally pure and their surface is so perfect that even a magnification of 40000 times fails to reveal any traces of roughness. Such perfection is mainly due to the condition of growth of small-size crystals, in which the freezing-in of lattice imperfections is less probable because it is easier for them to leave the crystal through a nearby surface.

Because of the absence of dislocations and of other defects in filament crystals a shift in a slip plane can only take the form of a rigid shift, in which the bonds of all the atoms in the slip plane are simultaneously broken. Stresses close to the theoretical stress limit of the crystals are needed to effect such a shift and this is what is observed in practice.

An unnaturally great elastic deformation of the whiskers is due to the absence of mobile dislocations, which in normal crystals are responsible for the plastic deformations occurring already at very low stresses.

Hence, the first method, the method of producing imperfection-free (in particular, dislocation-free) crystals, holds out a promise of producing materials of extreme strength close to the theoretical strength of solids.

(2) The second method is a direct opposite of the first. It consists in the maximum deformation of the internal structure of a crystal through the introduction of impurities, precipitation of dispersed phases, great plastic deformation, etc. Such defects hinder the motion of the dislocations and the growth of cracks and thus increase the strength of the material, as was already discussed in detail above. Science and industry have up to now made use only of this method and succeeded in attaining with it a strength of the order of 4×10^9 Pa. The effect this had on technology may be inferred from the following example. The specific weight of a modern aircraft engine is about 1 kgf per hp; at the turn of the century it was about 250 kgf per hp.

The recent times have witnessed the appearance of composite materials consisting of a matrix filled with filament crystals. Stainless steel, nickel, titanium and other materials are used for the matrix. The matrices are filled with tungsten, aluminium oxide, etc. filaments. The results obtained so far hold out a promise of obtaining by this method in the near future materials of 5 to 10 times the strength (especially at elevated temperatures) of the best steels and of 1.5-2 times lighter weight.

The strength of amorphous bodies and glass polymers is no less sensitive to internal structure. The strength of glass and quartz filaments newly extruded at a high temperature and practically free from defects² is 100 times as high as that of

²Since the atomic structure of amorphous bodies is irregular, the term defect may apply only to inclusions (clusters of foreign atoms, cracks, inhomogeneities) large if compared with atomic dimen-

normal specimens and quite close to the theoretical value.

The room temperature strength of unoriented glass polymers is of the order of 10^8 Pa. Films and fibres made of them having an oriented structure have a strength of the order of 10^9 Pa comparable to that of high quality steels. With a perfect orientation of the polymer molecule chain the strength of the needlelike crystals of the polymer may be as high as 3×10^{10} Pa. If one takes into account that the density of the polymers is close to unity, one can imagine how great their value for technology may be.

There is a rapidly growing demand on the quality of the materials for modern science and technology. Already now there is a need for materials able to withstand temperatures of several thousand degrees with the necessary strength characteristics at such temperatures and without any noticeable plastic deformation at normal loads.

What are the prospects for such materials?

One of the feasible methods for producing such extra strong and extra heat-resistant materials was proposed by the Soviet physicist A. V. Stepanov who pointed to a particular property of such molecular crystals as sulfur. The crystal of sulfur is constructed of molecules bonded by relatively weak molecular forces. Because of that the strength of the crystal and its melting point are low (115°C). The atoms in the sulfur molecule itself, on the other hand, are held together by powerful chemical bonds. If one would be able to construct a sulfur lattice with the atoms retaining the same bonds that act in the molecule, the result would be an extremely strong crystal with the melting point of about 34700°C . Similar modifications could be introduced into other molecular crystals as well. Are there any real grounds for such projects? The fact that we were able to transform soft graphite and hexagonal boron nitride into extra strong, hard, and high melting point diamond and borazon crystals by substituting powerful covalent bonds for weak van der Waals forces lends ground to such hopes. The prospects that will be opened by such materials are so enormous that any work, no matter how great, put into their production shall be generously rewarded.

Chapter 3

Elements of Physical Statistics

Every solid is a system, or an ensemble, consisting of an enormous number of microscopic particles. Such systems obey specific *statistical laws*, which are the subject of statistical physics, or physical statistics.

The present chapter deals briefly with the principal elements of physical statistics needed to describe the properties of solids.

§ 23. Methods used to describe the state of a macroscopic system

There are two methods of describing the state of a system consisting of a great number of microscopic particles, the *thermodynamic* and the *statistical* method. Let us discuss them.

Thermodynamic description of a system. In the thermodynamic approach to the description of the state of a system consisting of an enormous number of particles the latter is regarded as a macroscopic system, it being of no interest of what type of particles it consists. Such a system is termed a *thermodynamic system*.

A thermodynamic system may be either *closed* or *open*. A closed system does not interact in any way with the surroundings, and an open system can exchange heat and/or work with the surroundings.

The state of a system in which it can remain infinitely long is termed the *equilibrium state*. It is uniquely determined by a set of independent physical parameters, the *state parameters*. The principal state parameters are the *volume* of the system V , the *pressure* p , and the *temperature* T . However, often those parameters are inadequate for a complete characteristic of the system. For a system made up of several substances one has also to know their concentrations; for a system in an electric or a magnetic field the intensities of these fields should be specified; etc.

Any change in a thermodynamic system involving the variation of at least one

state parameter is termed a *thermodynamic process*.

The sum of all types of energy of a closed system is termed the *internal energy* (E) of the system. It is made up of the kinetic energy of the particles constituting the system, of the potential energy of the interaction between the particles, and of the internal energy of the particles themselves (which shall not be considered here since it is not subject to change in usual processes).

The internal energy is a *function of state* of the system. This means that there is one and only one definite value of internal energy that corresponds to each state no matter how the system arrived at this state.

Interacting with the surroundings a thermodynamic system may receive or reject some amounts ΔQ of heat, may perform work ΔA or have work performed on it. In all cases the variation in internal energy of the system, dE , should be equal to the difference in the amount of heat received from outside, ΔQ , and the work ΔA performed by the system against external forces:

$$dE = \Delta Q - \Delta A. \quad (3.1)$$

This is the first law of thermodynamics.

It should be pointed out that in contrast to the internal energy the work ΔA and the amount of heat ΔQ depend not only on the initial and the final states of the system but on the way the state is changed as well. Since

$$\Delta A = p dV \quad (3.2)$$

where dV is the variation of the volume of the system the pressure in which is p , we may write (3.1) in the form

$$dE = \Delta Q - p dV. \quad (3.3)$$

The second law of thermodynamics maintains that the amount of heat ΔQ received by the system in a reversible process results in the increase of the entropy of the system by

$$dS = \frac{\Delta Q}{T} \quad (3.4)$$

where T is the temperature at which the heat is received. Substituting ΔQ from (3.4) into (3.3), we obtain

$$dE = T dS - p dV. \quad (3.5)$$

It follows from (3.5) that the system's internal energy can be changed at the expense of work performed or heat exchanged.

However, the system's energy may also change with the change in the number of particles it contains, for every particle leaving the system takes away a definite amount of energy with it. Therefore, the general expression for the law of conser-

variation of energy (3.5) should be written in the form

$$dE = T dS - p dV + \mu dN \quad (3.6)$$

where dN is the variation of the number of particles in the system. Parameter μ is termed the *chemical potential* of the system. Its physical meaning is as follows. For an isolated system of constant volume which neither receives nor gives away heat, $dS = \Delta Q/T = 0$ and $dV = 0$. For such a system

$$dE = \mu dN. \quad (3.7)$$

Whence

$$\mu = \frac{dE}{dN}. \quad (3.8)$$

Hence, the chemical potential expresses the variation of the energy of an isolated system of a constant volume brought about by a unit variation in the number of particles it contains.

Let us consider the conditions of equilibrium of a system whose total number of particles remains constant but the particles can go over from one body belonging to the system to another. Two electron conductors, for instance, two metals, in contact with each other at a constant temperature may serve as an example of such a system. Denote the chemical potential of the electron gas in the first metal by μ_1 and in the second by μ_2 . Suppose dN electrons flow from one metal to another. According to (3.7) this will reduce the energy of the first metal by $dE_1 = \mu_1 dN$ and increase the energy of the second by $dE_2 = \mu_2 dN$. For the metals to be in a state of equilibrium the necessary condition is

$$dE_1 = dE_2, \quad \text{or} \quad \mu_1 dN = \mu_2 dN.$$

Hence the condition of equilibrium is

$$\mu_1 = \mu_2. \quad (3.9)$$

This condition is valid not only in the case of two electron conductors in contact with each other but for any phases in contact with each other: the solid and the liquid, the liquid and the gaseous, etc. *In all cases the condition of equilibrium is the equality of the chemical potentials.*

Statistical method of describing a system. To describe the state of every particle one should specify its three coordinates and three components of the momentum. Apparently, if one was to write the equations of motions of the particles and solve them, he would be able to obtain complete information on the behaviour of the system and to predict its state at any moment of time. Such calculations, however, are not only extremely tedious but, in fact, useless. The complexity of the problem stems from the fact that to describe the behaviour of the gas molecules normally contained in 1 m^3 one would have to solve about 10^{26} interconnected

equations of motion and also take into account the initial conditions, which is practically impossible. Should such calculations be carried out, they would be of no value since the properties of a system in the state of equilibrium not only are independent of the initial values of the coordinates and of the momentum components but generally remain constant in time, although the coordinates and the momenta of the particles do change. It follows from here that there is a qualitative distinction between the system and the individual particles and that the behaviour of the former is governed by laws different from those that govern the behaviour of individual particles. These laws are the statistical laws. The following examples are proof of their existence.

The velocity of an individual gas molecule is a random quantity, which is impossible to predict. Despite this fact, in a gas with a very large number of particles, on the average a distinct velocity distribution of its molecules may be observed. In other words, on the average a quite definite fraction of the molecules has a speed of, say, from 100 m s^{-1} to 200 m s^{-1} , from 400 m s^{-1} to 500 m s^{-1} , etc.

It is a matter of chance whether or not a given molecule shall enter a specified volume of the gas. Despite this fact there is a definite regularity in the distribution of the molecules over the volume: equal elements of volume contain, on the average, equal numbers of molecules.

The situation here is similar to that when a coin is tossed. The landing of the coin heads or tails up is a random event. Nevertheless, when the number of times the coin is tossed is very great, a quite definite regularity may be observed: on the average, the coin lands heads up half the number of times.

Such regularities are termed *statistical*. The principal feature of statistical laws is that they deal with *probabilities*. They enable predictions to be made only as to the probability of some event occurring or some result being realized. In the example with the coin the predicted probability of the coin landing one or the other side up is $1/2$. The results of individual tests may, and undoubtedly shall, deviate from those values the more the less the number of tests is. If we toss a coin five times, the head may fall out any number of times from 0 to 5. But the greater the number of tosses, that is, the more numerous the ensemble, the more accurate the statistical predictions are. Calculations show the relative deviation of an observed physical quantity (for instance, of the number of particles per unit volume) from the average value \bar{M} in a system of N noninteracting particles to be

$$\frac{\sqrt{\Delta M^2}}{\bar{M}} \propto \frac{1}{\sqrt{N}}$$

or inversely proportional to \sqrt{N} .

As N is increased, the ratio $\Delta M/\bar{M} \rightarrow 0$. For very great N we have that

$M/\overline{M} \approx 1$. Thus 1 m^3 of air normally contains on the average 2.7×10^{25} molecules. The relative deviation from this number is on the average equal to

$$\frac{100\%}{\sqrt{N}} \approx 2 \times 10^{-11}\%.$$

This deviation is so negligible that there are no instruments capable of detecting it. Therefore when dealing with large volumes it is always reasonable to assume that the distribution of molecules over the volume is uniform.

It should, however, be pointed out that deviations from the average values are not a possibility but a necessity. Such deviations are termed *fluctuations*.

§ 24. Degenerate and nondegenerate ensembles

Microscopic particles and the ensemble. All microscopic particles making up an ensemble may be subdivided into two classes according to their behaviour: fermions and bosons.

Fermions include electrons, protons, neutrons and other particles with a half-odd integral values of spin: $\hbar/2, 3\hbar/2, \dots$. Bosons include photons, phonons and other particles with integral values of spin: $0, \hbar, 2\hbar, \dots$

The fermions in an ensemble exhibit marked “individualistic” tendencies. If some quantum state is already occupied by a fermion, no other fermion shall settle in it. This is the essence of the well-known *Pauli exclusion principle*, which governs the behaviour of fermions. Bosons, on the other hand, strive for “unification”. They can settle in the same state in any numbers and do it the more readily the more populated the state already is.

Degenerate and nondegenerate ensembles. Let us discuss the possible effects of the nature of the particles (their fermionic or bosonic character) on the properties of the ensemble as a whole.

For the nature to be felt the particles must “meet” often enough. This means that they must occupy the same state or at least sufficiently closely-spaced states.

Suppose that there are G different states which any one of N similar particles can occupy. The ratio N/G may serve as a measure of the “meeting” frequency. The meetings will be rare if

$$\frac{N}{G} \gg 1. \quad (3.10)$$

In this case the number of different vacant states is much larger than the number of particles: $G \gg N$. Evidently, in such circumstances the specific nature of fermions and bosons shall not be felt, since every particle has at its disposal a large number of different free states and the problem of several particles occupying the same

state actually does not arise. Therefore, the properties of the ensemble as a whole shall not depend on the nature of the particles that make it up. Such ensembles are termed *nondegenerate*, and condition (3.10) is the *condition of nondegeneracy*.

If, however, the number of states G is of the same order of magnitude as the number of particles N , that is if

$$\frac{N}{G} \approx 1 \tag{3.11}$$

then the problem of how the states should be occupied, whether individually or collectively, indeed assumes much importance. In this case the nature of the microscopic particle is fully revealed in its effect on the properties of the ensemble as a whole. Such ensembles are termed *degenerate*.

Degenerate ensembles are a unique property of quantum objects, since the parameters of state of such objects only change discretely with the result that the number of possible states G can be finite. The number of states for classical objects whose parameters change continuously is always infinite and they can form only nondegenerate ensembles.

It should be pointed out that quantum mechanical objects too may form nondegenerate ensembles provided condition (3.10) is fulfilled (see Table 3.1).

Classical and quantum statistics. Physical statistics that studies nondegenerate ensembles is termed *classical statistics*. It owes much to J. C. Maxwell and L. E. Boltzmann (the *Maxwell-Boltzmann statistics*).

Physical statistics that studies degenerate ensembles is termed *quantum statistics*. Owing to the effect of the particles' nature on the properties of a degenerate ensemble, degenerate ensembles of fermions and bosons behave in essentially different ways. On this ground a distinction is made between two quantum statistics.

Quantum statistics of fermions owes much to E. Fermi and P. A. M. Dirac (this, by the way, explains the origin of the term "fermion"). It is termed the *Fermi-Dirac statistics*.

Quantum statistics of bosons owes much to S. N. Bose and A. Einstein (hence

Table 3.1

Object	Ensembles	
	Degenerate	Nondegenerate
Classical	No	Yes
Quantum	Yes	Yes

the term “boson”). It is termed the *Bose-Einstein statistics*.

It follows then that quantum statistics deals only with quantum objects while classical statistics may deal both with the classical and the quantum objects. If we reduce the number of particles in an ensemble or increase the number of states, we shall eventually turn a degenerate ensemble into a nondegenerate one. In that case the ensemble shall be described by the Maxwell-Boltzmann statistics no matter whether it contains fermions or bosons.

Distribution function. What is the connection between the distribution of the particles over particular states and the state of the ensemble as a whole? To specify the state of an ensemble, for instance, of a gas of particles, one should specify its state parameters. To specify the state of each particle one should specify its coordinates and momentum components or its energy, which is a function of coordinates and momentum.

The two types of quantities are connected by the *statistical distribution function*

$$N_{\mu,T}(E) dE \quad (3.12)$$

which specifies the number of particles having an energy from E to $E + dE$ in the system described by the state parameters μ and T .

This function is termed *complete statistical distribution function*. To simplify notation the indices denoting the state parameters are usually omitted.

The complete distribution function may be represented by the product of the number of states $g(E) dE$ per energy interval dE and the probability of occupation of those states by the particles. Let us denote the latter by $f(E)$. Then

$$N(E) dE = f(E)g(E) dE. \quad (3.13)$$

The function $f(E)$ is termed simply the *distribution function*. As was stated before it signifies the probability of the occupation of the respective states by the particles. If, for instance, 10 of 100 closely spaced states are occupied by particles (the total number of particles in the system being much greater than 100), the probability of occupation of such states will be equal to 0.1. Since on the average there is 0.1 of a particle per each state, we may take $f(E)$ to be the average number of particles in a given state.

Hence, the problem of finding the complete distribution function of particles over the states is reduced to that of finding the function $g(E) dE$, which describes the energy distribution of the states, and the function $f(E)$ which determines the probability of their occupation.

We start by determining the function $g(E) dE$.

§ 25. The number of states for microscopic particles

Concept of phase space of a microscopic particle and quantization. In classical mechanics the state of a particle is determined if its three coordinates (x, y, z) and three components of its momentum (p_x, p_y, p_z) are specified. Let us imagine a six-dimensional space with the coordinate axes x, y, z, p_x, p_y, p_z . The state of the particle at every moment of time will be described by a point (x, y, z, p_x, p_y, p_z). Such space is termed *phase space* and the points (x, y, z, p_x, p_y, p_z) are termed *phase points*. The quantity

$$\Delta\Gamma = \Delta\Gamma_V \Delta\Gamma_p = dx dy dz dp_x dp_y dp_z \quad (3.14)$$

is termed an element of the phase space. Here, $\Delta\Gamma_V = dx dy dz$ is an element of volume in coordinate space and $\Delta\Gamma_p = dp_x dp_y dp_z$ an element of volume in momentum space.

Since the coordinates and the momentum components of a classical particle may change continuously, the elements $\Delta\Gamma_V$, $\Delta\Gamma_p$ and with them the element $\Delta\Gamma$ as well can be chosen as small as desired.

The potential energy of a system of noninteracting particles not acted upon by an external field is zero. Such particles are termed *free*. For such particles it is convenient to use a three-dimensional momentum space instead of the six-dimensional phase space. In this case, the element $\Delta\Gamma_V$ is simply equal to the volume V in which the particles move, because no additional restrictions are placed on them.

The division of the phase space into elements of volume is not quite so simple if the particle in question is an electron or some other microscopic object possessing wave properties. The wave properties of such particles make it impossible, in accordance with the uncertainty principle, to distinguish between two states, (x, y, z, p_x, p_y, p_z) and ($x + dx, y + dy, z + dz, p_x + dp_x, p_y + dp_y, p_z + dz$), if the product $dx dy dz dp_x dp_y dp_z$ is less than h^3 . Since this product represents an element of volume in six-dimensional phase space, it follows from the above that different quantum states shall correspond to different elements of volume in this space only if the size of those elements is no less than h^3 . Therefore quantum statistics makes use of an elementary cell in the six-dimensional space with the volume

$$\Delta\Gamma = \Delta\Gamma_V \Delta\Gamma_p = h^3. \quad (3.15)$$

The element of the three-dimensional momentum space for free particles for which $\Delta\Gamma_V = V$ is

$$\Delta\Gamma_p = \frac{h^3}{V}. \quad (3.16)$$

Each element of this kind has corresponding to it a definite quantum state.

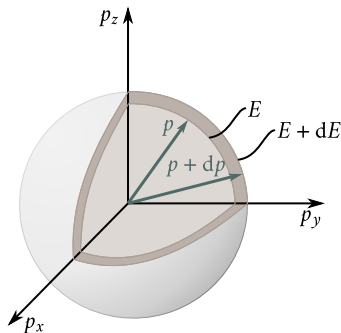


Figure 3.1: Calculating the number of states of a microscopic particle.

The process of dividing the phase space into cells of finite size (h^3 or h^3/V) is termed *quantization of phase space*.

Density of states. We wish to calculate the number of states of a free particle in the energy interval from E to $E + dE$. To this end draw two spheres of the radii p and $p + dp$ in the momentum space (Figure 3.1). There is a spherical layer with the volume of $4\pi p^2 dp$ contained between the spheres. The number of phase cells contained in this layer is

$$\frac{4\pi p^2 dp}{\Delta\Gamma_p} = \frac{4\pi V}{h^3} p^2 dp. \quad (3.17)$$

Since there is one particle state to correspond to every cell the number of states in the interval dp between p and $p + dp$ is

$$g(p) dp = \frac{4\pi V}{h^3} p^2 dp. \quad (3.18)$$

For free (noninteracting) particles

$$E = \frac{p^2}{2m}, \quad dE = \frac{p}{m} dp.$$

Using these relations to express p and dp and substituting the results into (3.18), we obtain

$$g(E) dE = \frac{2\pi V}{h^3} (2m)^{3/2} E^{1/2} dE. \quad (3.19)$$

This is the number of states of a free particle in the energy interval $(E, E + dE)$. Dividing the right- and the left-hand sides of (3.19) by dE , we obtain the density of states, $g(E)$, which specifies the number of states of a microscopic particle per unit energy interval:

$$g(E) = \frac{2\pi V}{h^3} (2m)^{3/2} E^{1/2}. \quad (3.20)$$

It follows from (3.20) that as E increases the density of states rises in proportion

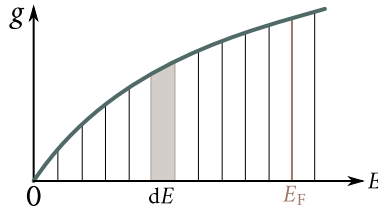


Figure 3.2: Energy dependence of density of states.

to $E^{1/2}$ (Figure 3.2). The density of states depends, besides, on the particle's mass and increases with m .

In case of the electrons each phase cell corresponds, to be exact, not to one but to two states, each distinguished by its spin. They are termed *spin states*. Therefore, in case of the electrons the number of states (3.18) and (3.19) and the density (3.20) should be doubled:

$$g(p) dp = \frac{8\pi V}{h^3} p^2 dp, \quad (3.21)$$

$$g(E) dE = \frac{4\pi V}{h^3} (2m)^{3/2} E^{1/2} dE, \quad (3.22)$$

$$g(E) = \frac{4\pi V}{h^3} (2m)^{3/2} E^{1/2}. \quad (3.23)$$

Condition of nondegeneracy for an ideal gas. Integrating (3.20) with respect to energy from 0 to E , we obtain the number of particle states contained within the energy interval $(0, E)$:

$$G = \frac{2\pi V}{h^3} (2m)^{3/2} \frac{2}{3} E^{3/2}.$$

Setting $E = 3k_B T/2$, we obtain

$$G \approx V \left(\frac{2\pi m k_B T}{h^2} \right)^{3/2}.$$

Substituting this expression into (3.10), we obtain the condition for nondegeneracy:

$$n \left(\frac{h^2}{2\pi m k_B T} \right)^{3/2} \ll 1 \quad (3.24)$$

where $n = N/V$ is the number of particles per unit volume.

Consider some molecular gas, for instance, nitrogen in normal conditions. For it $n \approx 10^{26} \text{ m}^{-3}$, $m = 4.5 \times 10^{-26} \text{ kg}$, and $k_B T = 4 \times 10^{-21} \text{ J}$. Substituting the figures into the left-hand side of (3.24), we obtain $nh^3(2\pi m k_B T)^{-3/2} \approx 10^{-6}$, which is much less than unity. Accordingly, the molecular gases are normally nondegenerate and must be described with the aid of the Maxwell-Boltzmann classical statistics.

Consider now the electron gas in metals. For it we have that $n \approx 5 \times 10^{24} \text{ m}^{-3}$ and $m = 9 \times 10^{-31} \text{ kg}$. For such values of n and m the electron gas turns out to be nondegenerate only at temperatures above 105 K; the left-hand side of (3.24) for such temperatures diminishes to less than unity (at $T = 105 \text{ K}$ it is approximately 0.5). Therefore, in practice the electron gas in metals is always degenerate and on account of this should be described with the aid of the Fermi-Dirac statistics.

It follows from (3.24) that a nondegenerate state of a gas can be realized not only by raising its temperature but by reducing its concentration n as well. For $n \approx 10^{22} \text{ m}^{-3}$ the left-hand side of (3.24) for electrons at normal temperatures is approximately 10^{-3} and the electron gas becomes nondegenerate. Such (and smaller) concentrations of the electron gas are found in some semiconductors. In such semiconductors termed *nondegenerate*, the electron gas is nondegenerate and is described by the classical Maxwell-Boltzmann statistics.

Let us try now to find the distribution function $f(E)$. The form of this function depends in the first instance on whether the gas is degenerate or nondegenerate. In the case of a degenerate gas the important point is whether the gas consists of fermions or bosons.

Let us start with a nondegenerate gas whose distribution function $f(E)$ is independent of the particles' nature.

§ 26. Distribution function for a nondegenerate gas

Appendix A.1 contains an elementary derivation of the distribution function for a nondegenerate gas. It is of the following form:

$$f(E) = e^{(\mu-E)/k_B T} \quad (3.25)$$

where k_B is Boltzmann's constant, and μ the chemical potential. Calculations give the following expression for μ of a nondegenerate gas:

$$\mu = k_B T \ln \left[\frac{N}{V} \left(\frac{h^2}{2\pi m k_B T} \right)^{3/2} \right]. \quad (3.26)$$

Substituting it into (3.25) we obtain:

$$f_M(E) = \frac{N}{V} \left(\frac{h^2}{2\pi m k_B T} \right)^{3/2} e^{-E/(k_B T)}. \quad (3.27)$$

We would like to remind again that $f_M(E) dE$ expresses the probability of occupation of the states in the energy interval $(E, E + dE)$; the term for it is the *Maxwell-Boltzmann distribution function*.

Figure Figure 3.3(a) shows a graph of the function $f_M(E)$. It has a maximum at $E = 0$ and asymptotically approaches zero as $E \rightarrow \infty$. This means that the lower

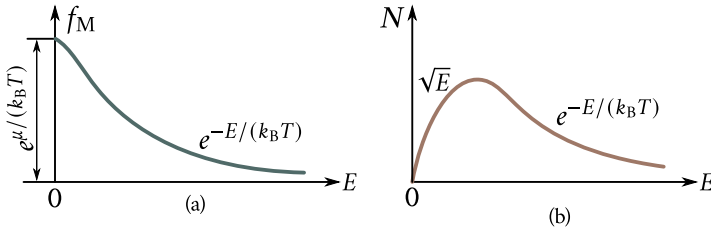


Figure 3.3: Distribution functions for nondegenerate gas: (a) —the Maxwell-Boltzmann distribution function expressing the average density of state occupation by particles; (b) — the complete Maxwell-Boltzmann distribution function.

energy states have the greatest probability of occupation. As the energy of a state increases its probability of occupation diminishes steadily.

Multiplying $f_M(E)$ by the number of states $g(E) dE$ [see Eq. (3.22)], we obtain the complete distribution function of particles over the energy

$$N(E) dE = \frac{4\pi V}{h^3} (2m)^{3/2} e^{\mu/(k_B T)} e^{-E/(k_B T)} E^{1/2} dE, \quad \text{or} \quad (3.28)$$

$$N(E) dE = \frac{2N}{\sqrt{\pi} (k_B T)^{3/2}} e^{-E/(k_B T)} E^{1/2} dE.$$

It is termed the *complete Maxwell-Boltzmann distribution function*. Figure 3.3(b) shows the graph of this function. Because of the factor $E^{1/2}$ its maximum is displaced to the right of the origin.

Knowing the distribution function $f_M(E)$ we may easily find the laws of distribution of the particles over the momentum, $N(p) dp$, and over its components, $N(p_x, p_y, p_z) dp_x dp_y dp_z$. We may also find them over the velocity, $N(v) dv$, and over its components $N(v_x, v_y, v_z) dv_x dv_y dv_z$ over one of the components of velocity, say $N(v_x) dv_x$, etc. Those distributions are shown below

$$N(p) = \frac{4\pi N}{(2\pi m k_B T)^{3/2}} e^{-p^2/(2m k_B T)} p^2, \quad (3.29)$$

$$N(v) = 4\pi N \left(\frac{m}{2\pi k_B T} \right)^{3/2} e^{-mv^2/(k_B T)} v^2, \quad (3.30)$$

$$N(p_x, p_y, p_z) = N \left(\frac{N}{2\pi m k_B T} \right)^{3/2} e^{-(p_x^2 + p_y^2 + p_z^2)/(2m k_B T)}, \quad (3.31)$$

$$N(v_x, v_y, v_z) = N \left(\frac{m}{2\pi k_B T} \right)^{3/2} e^{-m(v_x^2 + v_y^2 + v_z^2)/(k_B T)}, \quad (3.32)$$

$$N(v_x) = N \left(\frac{m}{2\pi k_B T} \right)^{1/2} e^{-mv_x^2/(k_B T)}. \quad (3.33)$$

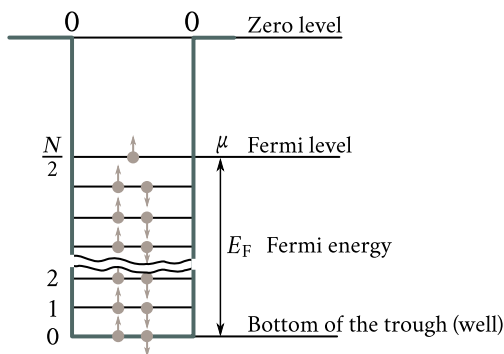


Figure 3.4: Schematic representation of a metal as a potential well for free electrons.

The reader is requested to obtain those results himself.

§ 27. Distribution function for a degenerate fermion gas

The distribution function for a degenerate fermion gas was first obtained by Fermi and Dirac. It is of the following form:

$$f_F(E) = \frac{1}{e^{(E-\mu)/(k_B T)} + 1}. \quad (3.34)$$

An elementary derivation of this expression is presented in Appendix A.2. Here, as before, μ , denotes the chemical potential, which in the case of a degenerate fermion gas is termed the *Fermi level*.

Equation (3.34) shows that for $E = \mu$, the distribution function $f(E) = 1/2$ at any temperature $T \neq 0$. Therefore, from the statistical point of view the Fermi level is a state whose probability of occupation is $1/2$.

The function (3.34) is termed the *Fermi-Dirac function*. To obtain a clear picture of the nature of this function one should consider the degenerate electron gas in metals at absolute zero.

Electron distribution in a metal at absolute zero. Fermi energy. The metal is a sort of a potential well (trough) for free electrons. To leave it the electron should have work performed on it against the forces retaining it in the metal. Figure 3.4 shows the diagram of such a potential well. The horizontal lines denote energy levels which the electrons may occupy. In compliance with the Pauli exclusion principle there may be two electrons with opposite spins on each such level. For an electron gas of N electrons the last occupied level will be the $N/2$ level. This level is termed the *Fermi level* for a degenerate electron gas. It corresponds to the maximum kinetic energy E_F an electron in a metal may possess at absolute zero. This energy is termed the *Fermi energy*.

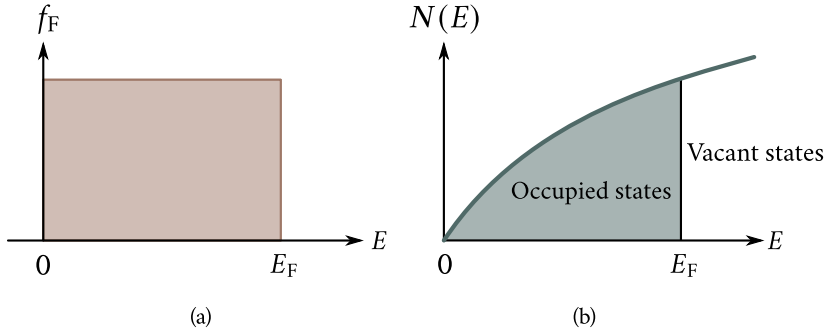


Figure 3.5: The distribution function for degenerate fermion gas: (a) — the Fermi-Dirac distribution function for $T = 0$ K; (b) — the complete Fermi-Dirac distribution function for $T = 0$ K.

Thus, at absolute zero all states with the energy $E < E_F$ are occupied by electrons and all states with the energy $E > E_F$ are free. In other words, the probability of occupation of a state with the energy $E < E_F$ at $T = 0$ K is unity and the probability of occupation of a state with the energy $E > E_F$ is zero:

$$f_F(E) = \begin{cases} 1, & \text{for } E < E_F, \\ 0, & \text{for } E > E_F. \end{cases} \quad (3.35)$$

To obtain this result from (3.34) one should assume that at $T = 0$ K the chemical potential of the electron gas measured from the bottom of the potential well is equal to the Fermi energy E_F :

$$\mu = E_F. \quad (3.36)$$

Indeed, setting in (3.34) $\mu = E_F$ we obtain

$$f_F(E) = \frac{1}{e^{(E-E_F)/(k_B T)} + 1}. \quad (3.37)$$

If $E < E_F$, then $e^{(E-E_F)/(k_B T)} \rightarrow 0$ at $T = 0$ K and $f_F = 1$; if $E > E_F$, then $e^{(E-E_F)/(k_B T)} \rightarrow \infty$ at $T = 0$ K and $f_F = 0$.

Figure 3.5(a) shows the graph of the Fermi-Dirac distribution function at absolute zero. It has the shape of a step terminating at $E = E_F$.

Multiplying (3.35) by the number of states $g(E) dE$ [see Eq. (3.22)], we obtain the *complete Fermi-Dirac distribution function at absolute zero*:

$$N(E) dE = \frac{4\pi V}{h^3} (2m)^{3/2} E^{1/2} dE \quad (3.38)$$

because $f_F = 1$ in the energy interval $(0, E_F)$. The graph of the function is presented in Figure 3.5(b), where the area of the occupied states is shaded.

Integrating expression (3.38) from 0 to E_F , we obtain

$$N = \frac{8\pi V}{3h^3} E_F^{3/2} (2m)^{3/2}$$

whence the Fermi energy may be easily obtained

$$E_F = \frac{h^2}{2m} \left(\frac{3n}{8\pi} \right)^{2/3} \quad (3.39)$$

where $n = N/V$ is the concentration of electron gas in the metal.

Knowing the energy distribution function of the electrons, we may easily calculate the average energy of the electrons at absolute zero, \bar{E}_0 :

$$\bar{E}_0 = \frac{3}{5} E_F = \frac{3h^2}{10m} \left(\frac{3n}{8\pi} \right)^{2/3}. \quad (3.40)$$

Lastly, knowing E_F and \bar{E}_0 , we can calculate the maximum velocity v_F and the effective velocity v_{eff} (corresponding to average energy) of free electrons in a metal at absolute zero:

$$v_F = \left(\frac{2E_F}{m} \right)^{1/2}, \quad v_{\text{eff}} = \left(\frac{2\bar{E}_0}{m} \right)^{1/2}. \quad (3.41)$$

Table 3.2 shows the Fermi energy E_F , the average energy \bar{E}_0 , the maximum and effective velocities of free electrons at absolute zero for some metals. The last column contains the Fermi temperature determined from the relation

$$T_F = \frac{E_F}{k_B}. \quad (3.42)$$

This is the temperature at which a molecule in a normal nondegenerate gas would have the energy of thermal motion $3k_B T/2$ equal to the Fermi energy E_F multiplied by $3/2$.

It may be seen from Table 3.2 that Fermi temperatures are so high that no metal can exist in a condensed state.

Table 3.2

Metal	E_F (eV)	\bar{E}_0 (eV)	v_F (10^6 m s $^{-1}$)	v_{eff} (10^6 m s $^{-1}$)	T_F (10^4 K)
Copper	7.10	4.30	1.60	1.25	8.20
Lithium	4.72	2.80	1.30	1.00	5.50
Silver	5.50	3.30	1.40	1.10	6.40
Sodium	3.12	1.90	1.10	0.85	3.70

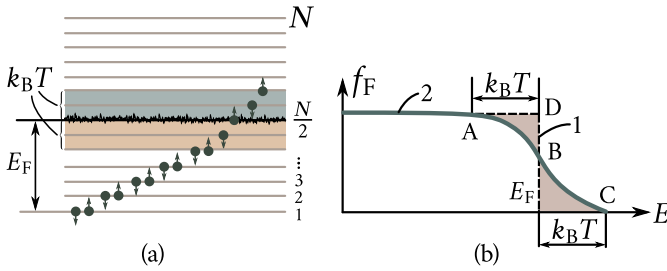


Figure 3.6: Temperature dependence of the Fermi-Dirac distribution function: (a) — thermal excitation of electrons; (b) — the Fermi-Dirac distribution function for $T > 0$ K.

It should be stressed that, although the Fermi energy represents the kinetic energy of translational motion of free electrons, it is not the energy of their thermal motion. Its nature is purely quantum mechanical and is due to the fact that electrons are fermions satisfying the Pauli exclusion principle.

Temperature dependence of the Fermi-Dirac distribution. When the temperature is raised, the electrons become thermally excited and go over to higher energy levels. This causes a change in their distribution over the states. However, in the range of temperatures in which the energy of their thermal motion, $k_B T$, remains much less than E_F only the electrons in a narrow band about approximately $k_B T$ wide adjoining the Fermi level may be thermally excited [Figure 3.6(a); the excited states are shaded]. The electrons of the lower levels remain practically unaffected because the energy of thermal excitation $k_B T$ is not enough to excite them (to transfer them to levels above the Fermi level).

As a result of thermal excitation some of the electrons with an energy less than E_F are transferred to the levels with energies greater than E_F and a new distribution of electrons over the states is established. Figure 3.6(b) shows the curves of the electron distribution over the states for $T = 0$ K (curve 1) and for $T > 0$ K (curve 2). It can be seen that the rise in temperature causes the original distribution to smear to a depth of $k_B T$ with a “tail” BC appearing to the right of E_F . The higher the temperature the greater the change in the distribution function. The tail BC itself is described by the Maxwellian distribution function.

The shaded areas in Figure 3.6(b) are proportional to the number of electrons transferred from the states with $E < E_F$ (the area ADB) to the states with $E > E_F$ (the area $E_F BC$). Those areas are equal since they represent the same number of electrons.

Let us make an approximate estimate of this number, ΔN . There are $N/2$ energy levels inside the interval $(0, E_F)$, where N is the number of free electrons in the metal. To simplify the problem we may assume that these levels are equidistant,

the separation being $\Delta\varepsilon = E_F/(N/2) = 2E_F/N$. Only the electrons in the band $k_B T$ wide just below E_F [Figure 3.6(a)] are thermally excited. There are $k_B T/\Delta\varepsilon = E_F N/(2E_F)$ levels inside this band occupied by $2k_B T N/(2E_F) = k_B T N/E_F$ electrons. Assuming that not more than a half of those electrons go over the Fermi level, we obtain the following approximate relation for ΔN :

$$\Delta N \approx \frac{k_B T}{2E_F} N. \quad (3.43)$$

At room temperature $k_B T \approx 0.025$ eV, $E_F = 3$ eV to 10 eV, therefore, $\Delta N/N < 1\%$; at $T = 1000$ K we find that $\Delta N/N \approx 1\%$ to 2% .

Hence, in all the temperature range in which the electron gas in a metal is degenerate its distribution is close to that at absolute zero.

Accordingly, only a negligible fraction of the electrons close to the Fermi level are thermally excited. At room temperature this fraction is less than 1% of the total number of conduction electrons. The laws governing the distribution of the electrons in metals discussed above remain valid practically in all cases, because in all the temperature range in which the existence of metals in the condensed state is possible the electron gas remains degenerate.

Consider the temperature dependence of the chemical potential. Integrating the complete Fermi-Dirac distribution function $f_F(E)g(E) dE$ over energy, we obtain the total number N of free electrons in a metal:

$$N = \int_0^\infty f_F(E)g(E) dE = \frac{4\pi V}{h^3} (2m)^{3/2} \int_0^\infty \frac{E^{1/2}}{[e^{(E-\mu)/(k_B T)} + 1]} dE.$$

Generally, there is no analytic expression for this integral. Approximate calculation in the temperature range in which the electron gas remains strongly degenerate yields the following temperature dependence of μ :

$$\mu = E_F \left[1 - \frac{\pi^2}{12} \left(\frac{k_B T}{E_F} \right)^2 \right]. \quad (3.44)$$

As $k_B T$ remains much less than E_F up to the melting point of a metal, the decrease in μ with the rise in T turns out to be so small that it can often be neglected and the Fermi level can be assumed to coincide with E_F at any temperature.

One can also calculate the average energy of the electrons \bar{E} in a degenerate electron gas dividing its total energy, $E_t = \int_0^\infty E f_F(E)g(E) dE$, by the number of the electrons, N :

$$\bar{E} = \frac{E_t}{N} = \frac{\int_0^\infty \frac{E^{3/2}}{[e^{(E-\mu)/(k_B T)} + 1]} dE}{\int_0^\infty \frac{E^{1/2}}{[e^{(E-\mu)/(k_B T)} + 1]} dE}.$$

An approximate calculation of these integrals yields

$$\bar{E} = \frac{3}{5}E_F \left[1 + \frac{5\pi^2}{12} \left(\frac{k_B T}{E_F} \right)^2 \right]. \quad (3.45)$$

At $T = 0$ K turns into (3.40).

Lifting of degeneracy. Nondegenerate electron gas. When the condition for nondegeneracy (3.10) is fulfilled, every gas including the electron gas must become nondegenerate. Let us discuss this in more detail.

According to (3.10) the gas is nondegenerate if the average occupancy of the states by the particles is much less than unity. Since the distribution function $f(E)$ represents the occupancy of the states, the condition of nondegeneracy (3.10) may be written in the form

$$f(E) \ll 1. \quad (3.46)$$

The Fermi-Dirac function (3.34) will be much less than unity if the exponential term, $e^{(E-\mu)/(k_B T)}$, will be much greater than unity:

$$e^{(E-\mu)/(k_B T)} \gg 1. \quad (3.46')$$

This inequality should hold for all the states including that with $E = 0$:

$$e^{(E-\mu)/(k_B T)} \gg 1. \quad (3.47)$$

It follows from (3.47) that for a nondegenerate electron gas in which condition (3.46) is satisfied, $-\mu$ should be a positive quantity considerably greater than $k_B T$:

$$-\mu \gg k_B T. \quad (3.48)$$

The chemical potential μ should be negative and greater than $k_B T$ in absolute value.

If the condition (3.46') is fulfilled, unity in the denominator of the Fermi-Dirac function can be neglected and the following expression for the distribution function of a nondegenerate electron gas may be obtained:

$$f(E) = e^{\mu/(k_B T)} e^{-E/(k_B T)}. \quad (3.49)$$

Comparing (3.49) with (3.25), we see that a nondegenerate electron gas like every other nondegenerate gas is described by the Maxwell-Boltzmann distribution function.

The electron gas in metals, where the free electron concentration is always very high ($\approx 10^{28} \text{ m}^{-3}$), is always in a degenerate state described by the Fermi-Dirac distribution function.

A nondegenerate electron gas is a feature of the *intrinsic (pure)* and weakly doped semiconductors, which are the mainstay of modern semiconductor electronics. The concentration of free electrons in such semiconductors is substantially less than in metals, varying from 10^{16} - 10^{19} m^{-3} to 10^{23} - 10^{24} m^{-3} depending upon the concentration of electrically active impurities. The nondegeneracy con-

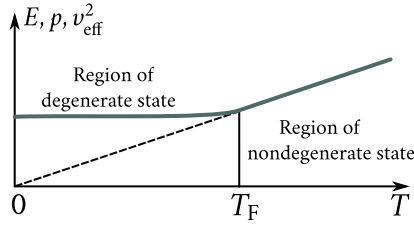


Figure 3.7: Temperature dependence of energy, pressure, and the square of effective velocity of electrons in a metal.

dition (3.10) remains valid for such concentrations and the electron gas is nondegenerate.

In conclusion, to illustrate the drastic difference in the behaviour of the ideal nondegenerate gas obeying the Maxwell-Boltzmann statistics and of the degenerate electron gas described by the Fermi-Dirac statistics, we would like to cite some of their properties (Table 3.3).

It follows that average energy \bar{E} , effective velocity v_{eff} and pressure p of a nondegenerate ideal gas are functions of temperature that vanish at absolute zero. At the same time \bar{E} , and p of a degenerate electron gas are very large already at absolute zero and are practically independent of temperature (Figure 3.7). This points to the fact that, as noted above, \bar{E} , v_{eff} , and p of a degenerate electron gas are for the most part not of thermal origin, the contribution of the thermal electron motion to these quantities being negligible.

Table 3.3

Parameters	Gas	
	Nondegenerate	Degenerate
\bar{E} at 0 K	0	$\bar{E}_0 = \frac{3}{5} E_F$
\bar{E} at T K	$\bar{E} = \frac{3}{2} k_B T$	$\bar{E} = \frac{3}{5} E_F \left[1 + \frac{5\pi^2}{12} \left(\frac{k_B T}{E_F} \right)^2 \right] \approx \frac{3}{5} E_F$
v_{eff} at 0 K	0	$v_{\text{eff},0} = \left(\frac{E_F}{m} \right)^{1/2} \approx 10^6 \text{ m s}^{-1}$
v_{eff} at T K	$v_{\text{eff}} = \left(\frac{3k_B T}{m} \right)^{1/2}$	$v_{\text{eff}} = v_{\text{eff},0}$
p at 0 K	0	$p_0 = \frac{2}{3} \bar{E}_0 \approx 10^{10} \text{ Pa}$
p at T K	$p = \frac{RT}{V}$	$p \approx p_0$

§ 28. Distribution function for a degenerate boson gas

In contrast to the electrons, which satisfy the Pauli exclusion principle, the bosons can occupy both the free states and the states already occupied by other bosons the more readily the greater the occupancy of the latter.

The distribution function of bosons over the states was first obtained by Bose and Einstein. It is of the following form:

$$f_{\text{Bose}}(E) = \frac{1}{[e^{(E-\mu)/(k_B T)} - 1]}. \quad (3.50)$$

(An elementary derivation of this expression is given in Appendix A.3). It is termed the *Bose-Einstein distribution function*. Let us use it to describe the state of a photon gas.

Suppose that a cavity inside a black body at a temperature T is filled with equilibrium thermal radiation. From the quantum mechanical point of view this radiation may be regarded as consisting of an enormous number of photons constituting a photon gas. The photon's spin is $s = 1$. Therefore, photons are bosons, which implies that the photon gas should satisfy the Bose-Einstein distribution.

The photons have some peculiarities as compared with other bosons, for instance, the helium nucleus, $\frac{4}{2}\text{He}$:

- (1) The rest mass of a photon is zero.
- (2) All photons move with the same speed equal to that of light, c , but can have different energy E and momentum p ; E and p depend on the photon's frequency:

$$E = h\nu = \hbar\omega, \quad p = \frac{h\nu}{c} = \frac{\hbar\omega}{c} \quad (3.51)$$

where $\hbar = h/2\pi$ and $\omega = 2\pi\nu$. It follows from (3.51) that

$$E = pc. \quad (3.52)$$

- (3) The photons do not collide with one another. Therefore, an equilibrium distribution in a photon gas can be established only in the presence of a body capable of absorbing and emitting photons. The walls of the cavity in which the radiation is contained may serve as an example of such a body. The transformation of a photon of one frequency into a photon of another frequency takes place in the processes of absorption and subsequent emission.
- (4) The photons may be generated (in the act of emission) and annihilated (in the act of absorption) in any numbers. Therefore, the number of photons in a photon gas does not remain fixed but depends on the state of the gas. For specified values of V and T the photon gas in a state of equilibrium contains so many photons N_0 as are needed for the energy of the gas to be at its min-

imum. This makes it possible to express the condition for the equilibrium of the photon gas in the form:

$$\left(\frac{dE}{dN} \right)_{V,T} = 0. \quad (3.53)$$

Since according to (3.8) $(dE/dN)_{V,T} = \omega$ the equilibrium condition (3.53) means that $\mu = 0$. Hence, the chemical potential of an equilibrium photon gas is zero.

For a nondegenerate gas the chemical potential is negative and has a relatively great absolute value. The fact that for the photon gas $\mu = 0$ means that such gas is always degenerate.

Setting $\mu = 0$ in (3.50), we obtain the distribution function for the photon gas:

$$f_P(E) = \frac{1}{[e^{E/(k_B T)} - 1]} = \frac{1}{[e^{(\hbar\omega)/(k_B T)} - 1]}. \quad (3.54)$$

This formula was first obtained by Max Planck and is termed the *Planck formula*. It represents the average fraction of photons having the energy $E = \hbar\omega$. Using this formula, we may easily formulate the law for the energy distribution in the spectrum of a black body. The following expression can be obtained for the energy density of the radiation of such a body:

$$\rho(\omega) = \frac{2\hbar}{\pi c^3} \left[\frac{\omega^3}{e^{(\hbar\omega)/(k_B T)} - 1} \right]. \quad (3.55)$$

The readers are requested to derive this formula themselves making use of (3.54) and (3.18).

§ 29. Rules for statistical averaging

As was already stated, to specify the state of an ensemble one should specify its state parameters. To specify the state of a particle one should specify the values of its coordinates and momentum components.

The problem of going over from the parameters of the individual particles to the state parameters characterizing the ensemble involves the problem of transition from the dynamical laws describing the behaviour of the individual particles to the statistical laws describing the behaviour of the ensemble. To effect such a transition it is necessary to perform the averaging of the characteristics of motion of the individual particles assuming the chances of all the particles belonging to the ensemble to be identical. The state parameters of the ensemble are expressed in terms of the averaged parameters of the individual particles belonging to the ensemble.

To make the rules of averaging apparent, let us consider an ensemble of N iden-

tical particles each of which is capable of assuming one of the discrete set of energy values: E_1, E_2, \dots, E_m . Choose an arbitrary moment of time and note the energy every particle has at that moment. We obtain as a result a set of numbers $N(E_i)$ expressing the number of particles having the energy E_i . To determine the average energy \bar{E} of the particles we add up the energies of all of them and divide the sum by the number of particles. The total number of particles is $N = \sum_{i=1}^m N(E_i)$, and the sum of their energies is $\sum_{i=1}^m E_i N(E_i)$. Therefore,

$$\bar{E} = \frac{\sum_{i=1}^m E_i N(E_i)}{\sum_{i=1}^m N(E_i)}. \quad (3.56)$$

The average value E obtained in this fashion is termed an *ensemble average*.

If the particle's energy assumes a continuous set of values, the practice is to count the number of particles having an energy lying within an interval $(E, E + dE)$ instead of the number of particles having an exact value of energy. The average energy will then be

$$\bar{E} = \frac{\int_0^{\infty} E N(E) dE}{\int_0^{\infty} N(E) dE}. \quad (3.57)$$

Such averaging may be performed for any physical quantity M that is a function of the coordinates and the momenta of the particles making up the ensemble. If M is continuous,

$$\bar{M} = \frac{\int_0^{\infty} M N(M) dM}{\int_0^{\infty} N(M) dM}. \quad (3.58)$$

Let us determine the average energy of the particles of an ideal nondegenerate gas. According to (3.57) and (3.28), we have

$$\bar{E} = \frac{\int_0^{\infty} E N(E) dE}{\int_0^{\infty} N(E) dE} = \frac{2}{\sqrt{\pi} (k_B T)^3} \int_0^{\infty} e^{-E/(k_B T)} E^{3/2} dE = \frac{3}{2} k_B T. \quad (3.59)$$

The results of the calculations of the average values of the velocity, of a velocity component, of the effective velocity, and of its component for the particles of an

ideal gas are presented thus:

$$\bar{v} = \left(\frac{8k_{\text{B}}T}{\pi m} \right)^{1/2}, \quad \bar{v}_x = \left(\frac{k_{\text{B}}T}{2\pi m} \right)^{1/2},$$
$$v_{\text{eff}} = \left(\frac{3k_{\text{B}}T}{m} \right)^{1/2}, \quad v_{\text{eff},x} = \left(\frac{k_{\text{B}}T}{m} \right)^{1/2}.$$

The reader may for the sake of practice perform these calculations himself.

Chapter 4

Thermal Properties of Solids

§ 30. Normal modes of a lattice

The atoms of a solid take part in thermal vibrations around their equilibrium positions. Because of a strong interaction between them, the nature of those vibrations turns out to be extremely complex and an accurate description of it presents enormous difficulties. Therefore, approximate methods and various simplifications are used to solve this problem.

Instead of describing the individual vibrations of the particles the practice is to consider their collective motion in the crystal, which is a spatially ordered structure. This simplification is based on the fact that powerful bonds immediately transmit the vibrations of one particle to other particles and a collective motion in the form of an elastic wave involving all the particles of the crystal is excited in it. Such collective motion is called the *normal mode of a lattice*. The number of normal modes coincides with the number of degrees of freedom, which is $3N$ if N is the number of particles constituting the crystal.

Figure 4.1(a) represents a one-dimensional model of a solid—a linear chain of atoms separated by a distance a and able to vibrate in the direction perpendicular to the chain. Such a chain may be regarded as a string. If the ends of the chain are fixed, the fundamental mode corresponding to the lowest frequency ω_{\min} is represented by the standing wave with a node at each end [Figure 4.1(b), curve 1]. The second mode is represented by the standing wave with an additional node in the centre of the chain (curve 2). The third mode, or third harmonic, is represented by the standing wave with two additional nodes that divide the chain in three equal parts (curve 3), etc. The length of the shortest wave in such a chain is evidently equal to twice the distance between the atoms of the chain [Figure 4.1(c)]:

$$\lambda_{\min} = 2a. \quad (4.1)$$

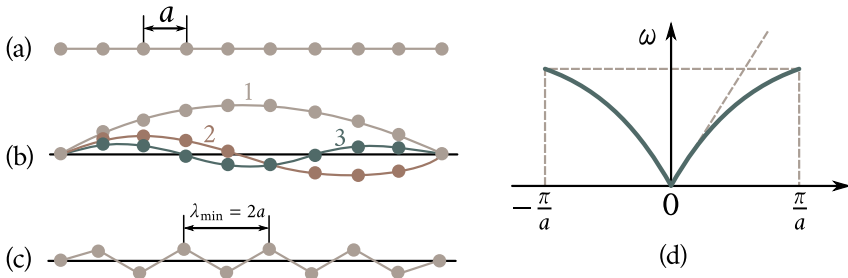


Figure 4.1: Normal modes of a linear chain made up of identical atoms: (a)—linear chain; (b)—normal modes of the chain; (c)—normal modes of the chain corresponding to shortest wavelength (to highest frequency); (d)—dispersion curves expressing dependence of normal mode frequency on wave vector.

The corresponding maximum frequency ω_{\max} is

$$\lambda_{\max} = \frac{2\pi v}{\lambda_{\min}} = \frac{\pi v}{a} \quad (4.2)$$

where v is the velocity of wave propagation (of sound) along the chain.

This maximum frequency is a parameter of the chain's material and is determined by the interatomic distance and the velocity of wave propagation. Should we set $a = 3.6 \times 10^{-10}$ m (the lattice parameter of copper) and $v = 3550$ m s⁻¹ (the velocity of sound in copper) we would obtain $\omega_{\max} \approx 3 \times 10^{13}$ s⁻¹, which corresponds to the frequency of atomic vibrations in a solid.

To describe wave processes one usually uses the wave vector \mathbf{q} whose direction coincides with that of wave propagation and whose absolute value is

$$q = \frac{2\pi}{\lambda}. \quad (4.3)$$

It follows from Eq. (4.2) that $2\pi/\lambda = \omega/v$. Therefore¹,

$$q = \frac{\omega}{v} \quad \Rightarrow \quad \omega = qv. \quad (4.4)$$

¹The phase velocity v , which enters (4.4), is itself a function of the wave vector \mathbf{q} and for a linear chain of atoms bonded by elastic forces is expressed by the following relation:

$$v = v_0 \frac{\sin(qa/2)}{(qa/2)} \quad (4.4')$$

where v_0 is the velocity of wave propagation in a continuous string for which $a = 0$. It follows from Eq. (4.4') that for a constant a , the velocity v is practically independent of \mathbf{q} and is approximately v_0 only in the range of small q 's, where $[\sin(qa/2)]/(qa/2) \ll 1$. In this range ω increases approximately in proportion to q [Figure 4.1(d)]. As q increases, the value of $[\sin(qa/2)]/(qa/2) \ll 1$ steadily diminishes and for $q = \pi/a$ tends to $2/\pi$. This causes the dispersion curve $\omega(q)$ to flatten out, so that for $q = \pi/a$ it runs parallel to q .

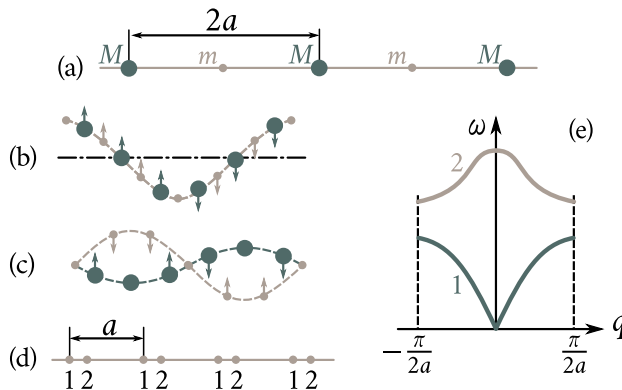


Figure 4.2: Normal modes of a chain made up of atoms of two kinds: (a)—arrangement of atoms in the chain; (b)—acoustic modes; (c)—optical modes; (d)—dispersion curves for acoustic and optical modes; (e)—linear lattice with a basis in which optical and acoustic modes are present.

Figure 4.1(d) shows the dependence on wave vector q of the frequency of normal modes in a linear chain made up of atoms of one kind. As q increases from 0 to π/a , the frequency of the normal modes rises, reaching the maximum value $\omega_{\max} = \pi v/a$ for $q = \pi/a$, that is, for $\lambda = 2a$. Curves of this type expressing the dependence of the vibration frequency on the wave vector (the wavelength) are termed *dispersion curves*.

Consider now a chain made up of atoms of two kinds placed in a regular sequence one after another [Figure 4.2(a)]. Denote the mass of the heavier atoms by M and that of the lighter atoms by m . Two types of normal modes can be present in such a chain, as is shown in Figure 4.2(b,c). The modes in Figure 4.2(b) are quite identical to the modes of a uniform chain: the phases of the neighbouring atoms are practically the same. Such vibrations are termed *acoustic modes*, since they include the entire spectrum of the acoustic modes of the chain. For them, $\omega_{\text{ac}} = 0$ for $q = 0$. They play a decisive part in determining the thermal properties of the crystals such as heat capacity, heat conductivity, thermal expansion, etc.

In case of normal modes shown in Figure 4.2(c) the phases of the neighbouring vibrating atoms are opposite. Such modes can be treated as the relative vibrations of two interpenetrating sublattices, each made up of atoms of one kind. They are termed *optical modes*, since they play a decisive part in the processes of interaction of light with crystals.

Figure 4.2(d) shows the dispersion curves for the acoustic (1) and optical (2) normal modes of a chain made up of atoms of two kinds. In contrast to an acoustic mode, whose frequency rises with the wave vector reaching the maximum value at

$q_{\max} = \pi/(2a)$, the maximum frequency of an optical mode corresponds to $q = 0$; with the increase in q the frequency of an optical mode falls off to its minimum at $q_{\max} = \pi/(2a)$.

The optical vibrations are possible not only in a chain made up of atoms of different kinds, but in a complex chain made up of two, or more, interpenetrating chains containing atoms of one kind, as shown in Figure 4.2(e). Unit cell of such a complex chain contains two atoms. The optical modes are the result of the relative vibrations of two sublattices.

§ 31. Normal modes spectrum of a lattice

One of the principal problems of the theory of lattice vibrations is the problem of the frequency distribution of normal modes. Consider now the simplest case of the normal modes in a linear atomic chain (see Figure 4.1).

The wavelengths of the normal modes in such a chain are

$$\lambda_n = \frac{2L}{n} \quad (n = 1, 2, 3, \dots, N) \quad (4.5)$$

where L is the length of the chain, and N the number of atoms in it.

The number of normal modes z having the wavelength equal to or greater than λ_n will evidently be n :

$$z = n = \frac{2L}{\lambda_n}.$$

In the same way the number of standing waves in a three-dimensional crystal of volume V (for instance, in a cube with edge L and volume L^3) having the wavelength equal to or greater than λ should be

$$z = \left(\frac{2L}{\lambda} \right)^3 = \frac{8V}{\lambda^3}.$$

A more accurate calculation yields

$$z = \frac{4\pi V}{\lambda^3}. \quad (4.6)$$

Since $\lambda = 2\pi v/\omega$, it follows that

$$z = \frac{V}{2\pi^2 v^3} \omega^3. \quad (4.7)$$

Differentiating this expression, we obtain

$$dz = g(\omega) d\omega = \frac{3V}{2\pi^2 v^3} \omega^2 d\omega. \quad (4.8)$$

Equation (4.8) expresses the number of normal modes per frequency interval

$(\omega, \omega + d\omega)$. The function

$$g(\omega) = \frac{dz}{d\omega} = \frac{3V}{2\pi^2 v^3} \omega^2 \quad (4.9)$$

determines the *density* of the normal vibrations in $d\omega$ of the spectrum, that is their spectrum. The function $g(\omega)$ is termed *spectral distribution function of normal modes*.

Since the number of normal vibrations in a lattice is $3N$, the function $g(\omega)$ should satisfy the following normalization condition:

$$\int_0^{\omega_D} g(\omega) d\omega = 3N \quad (4.10)$$

where ω_D is the maximum frequency limiting the spectrum of normal modes from above.

Substituting (4.9) into (4.10) and integrating, we obtain

$$\frac{V\omega_D^3}{2\pi^2 v^3} = 3N. \quad (4.11)$$

Hence,

$$\omega_D = v \left(6\pi^2 \frac{N}{V} \right)^{1/3}. \quad (4.12)$$

The frequency ω_D is termed the *Debye frequency* and the temperature

$$\Theta = \frac{\hbar\omega_D}{k_B} \quad (4.13)$$

the *Debye temperature*. Table 4.1 shows the Debye temperatures of some chemical elements and compounds.

At the Debye temperature the entire vibration spectrum is excited in the solid, including the one with the maximum frequency ω_D . Accordingly, any further rise in temperature (above Θ) shall not be accompanied by the appearance of new normal modes. In this case, the role of the temperature is to increase the intensity of each of the normal modes with a resultant increase in their average energy.

Temperatures $T > \Theta$ are usually referred to as *high temperatures*.

Substituting v^3 from (4.11) into (4.9), we obtain

$$g(\omega) = 9N \frac{\omega^2}{\omega_D^3}. \quad (4.14)$$

§ 32. Phonons

Each normal mode carries with it some energy and momentum. Oscillation theory contains the proof of the fact that the energy of a normal mode is equal to the

energy of an oscillator with a mass equal to the mass of the vibrating atoms and the frequency of the normal mode. Such oscillators are *termed normal*.

Denote the energy of the i -th mode characterized by the frequency ω_i by $E_{i,n,m}$. It is equal to the energy $E_{i,n,o}$ of a normal oscillator of the same frequency ω_i : $E_{i,n,m} = E_{i,n,o}$. The total energy of the crystal in which all the $3N$ normal modes have been excited is

$$E = \sum_i^{3N} E_{i,n,o}.$$

Hence, the total energy of the crystal made up of N atoms taking part in coupled vibrations is equal to the energy of $3N$ independent normal harmonic linear oscillators. In this sense, a system of N atoms whose vibrations are interconnected is equivalent to a set of $3N$ normal oscillators and the problem of calculating the average energy of such a system is reduced to a simpler problem of calculating the average energy of normal oscillators.

It should be pointed out that normal oscillators have nothing in common with real atoms except the mass. Every oscillator represents one of the normal modes of the lattice in which all the atoms of the crystal take part vibrating with the same

Table 4.1

Element	Θ , (K)	Element	Θ , (K)	Element	Θ , (K)
Al	418	Fe	467	Pb	94.5
Ag	225	Ge	366	Pd	275
Au	165	Hg	275	Pt	229
Be	1160	In	109	Si	658
Bi	117	KBr	174	Sn (gray)	212
C (diamond)	1910	KCl	227	Sn (white)	189
Ca	219	La	132	Ta	231
CaF ₂	474	Mg	406	Ti	278
Cd	300	Mo	425	Tl	89
Co	445	NaCl	320	V	273
Cr	402	Nb	252	W	(379)
Cu	339	Ni	456	Zn	308

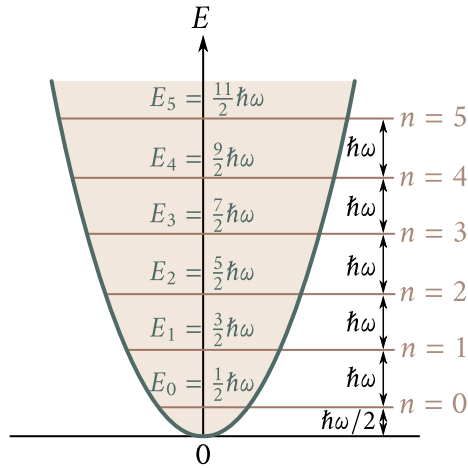


Figure 4.3: Energy spectrum of linear harmonic oscillator.

frequency ω .

The energy of a quantum oscillator is, as is well known, expressed by the relation

$$E_n = \left(n + \frac{1}{2}\right) \hbar\omega \quad (n = 0, 1, 2, \dots) \quad (4.15)$$

where ω is the oscillator's vibration frequency, and n the quantum number.

Figure 4.3 shows the energy spectrum of a linear harmonic oscillator. It consists of a set of discrete levels spaced at an interval of $\hbar\omega$.

Since $E_{n,m} = E_{n,0}$, the expression for the energy of the normal modes of a lattice should be (4.15) and the energy spectrum should coincide with that shown in Figure 4.3.

The minimum portion of energy that can be absorbed or emitted by the lattice in the process of thermal vibrations corresponds to the transition of the normal mode being excited from the given energy level to the adjacent level and is equal to

$$\varepsilon_{ph} = \hbar\omega. \quad (4.16)$$

This portion, or quantum, of energy of thermal vibrations of the lattice is termed a *phonon*.

The following analogy may help to clear up the point. The space inside a black body is filled with equilibrium thermal radiation. From the quantum mechanical point of view such radiation is treated as a gas made up of the light quanta, or photons, whose energy is $\varepsilon = \hbar\omega = h\nu$ and whose momentum is $p = \hbar\omega/c = h/\lambda$, where c is the velocity of light, and λ its wavelength.

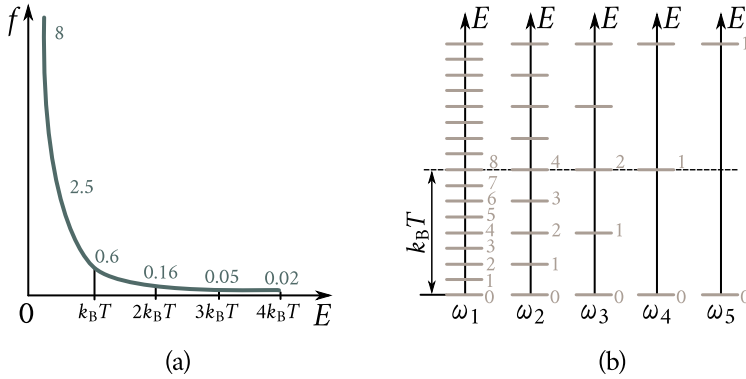


Figure 4.4: Energy spectrum of linear harmonic oscillator.

The field of elastic waves in the crystal may be treated similarly as a gas made up of quanta of the normal modes of the lattice, or of phonons having the energy $\varepsilon_{\text{ph}} = \hbar\omega = h\nu$ and momentum

$$p_{\text{ph}} = \frac{\hbar\omega}{v} = \frac{h}{\lambda} = \hbar q \quad (4.17)$$

where v is the velocity of sound, and λ the length of the elastic wave.

From this point of view a heated crystal may be likened to a box filled with phonon gas. The analogy may be extended.

Phonons are described by the same Bose-Einstein distribution function (3.54) as photons:

$$f(E) = \frac{1}{e^{\varepsilon_{\text{ph}}/(k_B T)} - 1} = \frac{1}{e^{(\hbar\omega)/(k_B T)} - 1}.$$

Depending on the intensity of excitation of the normal mode it can “emit” a definite number of phonons. Thus, if some normal mode was excited to the third level (Figure 4.3), its energy became $E_3 = (3 + 1/2)\hbar\omega$; this means that the particular normal mode has “generated” three identical phonons each with an energy of $\hbar\omega$.

Figure 4.4(a) shows the graph of the phonon energy (frequency) distribution function $f(E)$. We see that for a given temperature T , all normal modes in a lattice up to those with the energy $\hbar\omega \approx k_B T$ are excited; practically no quanta of higher frequencies with the energy $\hbar\omega > k_B T$ are excited. This is quite evident from Figure 4.4(b). Horizontal strokes here denote energy spectra of normal modes with the frequencies $\omega_1 = k_B T/(8\hbar)$, $\omega_2 = k_B T/(4\hbar)$, $\omega_3 = k_B T/(2\hbar)$, $\omega_4 = k_B T/\hbar$ and $\omega_5 = 2k_B T/\hbar$; the level corresponding to $k_B T$ is shown by a dotted line. It follows that for a given temperature T the mode with the frequency ω_1 is excited approximately to the 8th level. As was stated before, this means that this normal mode “generates” eight identical phonons with the energy $\hbar\omega_1 = k_B T$ each. The normal

mode with the frequency ω_2 is excited approximately to the 4th level, that with the frequency ω_3 to the second, and that with the frequency ω_4 (whose quantum of energy is $\hbar\omega_4 = k_B T$) to the first. At the same time, the vibration ω_5 is rarely excited at T because its excitation energy $\hbar\omega_5$ is too high. The excitation of still higher frequencies is a much more rare event. Therefore, we can say that approximately only the vibrations with frequencies not greater than ω corresponding to the energy $\hbar\omega \approx k_B T$ are excited in a solid at temperatures $T < \Theta$.

By definition, the distribution function $f(E)$ expresses the average number of phonons having the energy $\varepsilon_{\text{ph}} = \hbar\omega$. Therefore, to obtain the average energy of an excited normal mode, $\bar{E}_{\text{n.m.}}$, of the frequency ω one has to multiply (3.54) by $\hbar\omega$:

$$\bar{E}_{\text{n.m.}} = \frac{\hbar\omega}{e^{(\hbar\omega)/(k_B T)} - 1}. \quad (4.18)$$

§ 33. Heat capacity of solids

The thermal energy of a solid E_{lattice} is the sum of the energies of its normal modes. The number of normal modes per spectral interval $d\omega$ is $g(\omega) d\omega$ [see Eq. (4.8)]. Multiplying this number by the average energy $\bar{E}_{\text{n.m.}}$ of the normal mode, (4.18), we obtain the total energy of the normal modes in the interval $d\omega$

$$dE_{\text{lattice}} = \bar{E}_{\text{n.m.}} g(\omega) d\omega.$$

Integrating this expression over the entire spectrum of the normal modes, that is, from 0 to ω_D , we obtain the energy of the thermal vibrations of the lattice of a solid:

$$E_{\text{lattice}} = \int_0^{\omega_D} \bar{E}_{\text{n.m.}} g(\omega) d\omega. \quad (4.19)$$

The heat capacity at constant volume of a solid, C_V , is the change in the thermal energy of a solid brought about by a one degree change in its temperature. To find it one should differentiate lattice with respect to T :

$$C_V = \frac{dE_{\text{lattice}}}{dT}. \quad (4.20)$$

The fundamental problem in the theory of heat capacity is the temperature dependence of C_V . Let us first consider it from a qualitative point of view for two temperature ranges: for the range of temperatures much below the Debye temperature

$$T \ll \Theta \quad (4.21)$$

which is termed the *low temperature range*, and for the range of temperatures above the Debye temperature

$$T > \Theta \quad (4.22)$$

the term for which is the *high temperature range*.

Low temperature range. In this range mainly the low frequency normal modes, with the energy quanta $\hbar\omega < k_B T$, are excited. The approximate value of the average energy of normal vibrations may in this case be calculated with the aid of the following method. Expand the denominator of expression (4.18) into a series leaving only two terms:

$$\bar{E}_{n.m} = \frac{\hbar\omega}{e^{(\hbar\omega)/(k_B T)} - 1} \approx \frac{\hbar\omega}{1 + \frac{\hbar\omega}{k_B T} + \dots - 1}$$

Hence, in the low temperature range the average energy of every normal mode increases in proportion to the absolute temperature T :

$$\bar{E}_{n.m} \propto T. \quad (4.23)$$

This law is due to the increase in the probability of excitation of every normal mode with the rise in temperature resulting in an increase in its average energy.

In addition to this, the rise in temperature in the low temperature range, causes new higher frequency normal modes to be excited. The approximate number of the latter, z , may be calculated with the aid of Eq. (4.8). If we assume that at a temperature T all normal modes up to the frequency $\omega \approx k_B T / \hbar$ are excited, we get

$$z = \int_0^{k_B T / \hbar} g(\omega) d\omega \approx \int_0^{k_B T / \hbar} \omega^2 d\omega \propto T^3.$$

It follows that with the rise in temperature the number of normal modes increases in proportion to the cube of the absolute temperature:

$$z \propto T^3. \quad (4.24)$$

To sum up, the crystal's energy in the low temperature range increases with the rise in temperature by means of two mechanisms: (1) the increase in the average energy of every normal mode, $\bar{E}_{n.m}$, due to the rise in the probability of its excitation, and (2) the increase in the number of the normal modes of the lattice.

The first mechanism is responsible for the increase in energy proportional to T and the second for the one proportional to T^3 .

Therefore, the total effect is an increase in the energy of the lattice proportional to T^4 :

$$\bar{E}_{\text{lattice}} \propto T^4 \quad (4.25)$$

and a rise in heat capacity proportional to T^3 :

$$C_V \propto T^3. \quad (4.26)$$

Formula (4.26) is the *Debye T^3 law*, which agrees well with experiment in the

low temperature range.

High temperature range. As has been already stated, all normal modes of a lattice are excited at the Debye temperature, so that a further rise in temperature cannot increase their number. Therefore, the variation in energy of a solid in the high temperature range may only be due to the rise in intensity of the normal modes, resulting in an increase in their average energy $\bar{E}_{n.m.}$. Since $\bar{E}_{n.m.} \propto T$, the variation of the energy of the body as a whole, too, should be proportional to T :

$$\bar{E}_{\text{lattice}} \propto T \quad (4.27)$$

and the heat capacity must be independent of T :

$$C_V = \frac{d\bar{E}_{\text{lattice}}}{dT} = \text{constant}. \quad (4.28)$$

Relation (4.28) is the expression of the *Dulong and Petit law*, which is quite well substantiated by experiment.

A rather wide range of temperatures, the so-called *medium temperature range*, lies between the high and low temperature ranges. In this medium temperature range, gradual transition from the Debye T^3 law to the Dulong and Petit law takes place. Calculations in this range are the most difficult.

To sum up, the following physical picture of the variation of the temperature dependence of energy and of heat capacity of a solid with the rise in temperature may be presented.

In the low temperature range ($T \ll \Theta$) the solid's energy increases with the rise in temperature firstly because of the increase in the probability of excitation of every normal mode, that is because of the increase in its average energy, $\bar{E}_{n.m.}$, which is proportional to T , and secondly because new normal modes are drawn into the process causing the body's energy to increase in proportion to T^3 . The energy of the lattice, as a whole, rises in proportion to T^4 and the specific heat in proportion to T^3 .

As the temperature approaches the Debye temperature the second mechanism gradually becomes inoperative and E_{lattice} becomes less dependent on T , causing a deviation from the Debye T^3 law.

At the Debye temperature, the entire spectrum of normal modes is excited. Therefore, the second mechanism has no part to play; only the first mechanism operates here causing the energy to rise in proportion to T and the heat capacity C_V to remain independent of T (the Dulong and Petit law).

The qualitative laws of the variation of C_V with T obtained from the study of physical processes in solids may be substantiated by more rigorous quantitative calculations. To this end let us turn to (4.19) and try to calculate the lattice energy as a function of temperature more accurately.

Substituting $g(\omega)$ from Eq. (4.14), and $\bar{E}_{n.m}$ from (4.18) into (4.19), we obtain

$$E_{\text{lattice}} = \frac{9N}{\omega_D^3} \int_0^{\omega_D} \frac{(\hbar\omega)^3}{e^{(\hbar\omega)/(k_B T)} - 1} d\omega. \quad (4.29)$$

One can introduce the dimensionless quantity $x = (\hbar\omega)/(k_B T)$ and rewrite (4.29) in the form

$$E_{\text{lattice}} = 9Nk_B \Theta \left(\frac{T}{\Theta} \right)^4 \int_0^{\Theta/T} \frac{x^3}{e^x - 1} dx \quad (4.30)$$

where Θ is the Debye temperature.

We will consider the high and the low temperature ranges separately.

Low temperature range ($T \ll \Theta$). In this range we can substitute infinity for the limit of integration in (4.30). Taking into account that

$$\frac{x^3}{e^x - 1} dx = \frac{\pi^4}{15}$$

we obtain

$$E_{\text{lattice}} = \frac{3\pi^4}{5} Nk_B \Theta \left(\frac{T}{\Theta} \right)^4 \propto T^4. \quad (4.31)$$

Differentiating Eq. (4.31) with respect to temperature, we obtain

$$C_V = \frac{12\pi^4}{5} Nk_B \left(\frac{T}{\Theta} \right)^3 \propto T^3. \quad (4.32)$$

We have arrived at the Debye T^3 law in accordance with which the heat capacity of a lattice varies in the low temperature range as the cube of the temperature.

High temperature range. For such temperatures the values of x are small and, hence, it is possible to drop all but the first two terms of the expansion $e^x = 1 + x + \dots$. Then,

$$E_{\text{lattice}} = 9Nk_B \Theta \left(\frac{T}{\Theta} \right)^4 \int_0^{\Theta/T} x^2 dx = 3Nk_B T \propto T. \quad (4.33)$$

The heat capacity of the crystal is

$$C_V = \frac{dE_{\text{lattice}}}{dT} = 3Nk_B = \text{constant}. \quad (4.34)$$

For a mole of a monatomic substance $N = N_A = 6.023 \times 10^{23} \text{ mol}^{-1}$ (Avogadro's number), $N_A k_B = R \approx 8.31 \text{ J mol}^{-1} \text{ K}^{-1}$ (the gas constant) and

$$C_V \approx 3R \approx 25 \text{ J mol}^{-1} \text{ K}^{-1}. \quad (4.35)$$

Formula (4.35) expresses the Dulong and Petit law, which was formulated by them in 1819.

The solid line in Figure 4.5 shows the theoretical temperature dependence of the heat capacity of solids, the points being experimental values for silver, diamond,

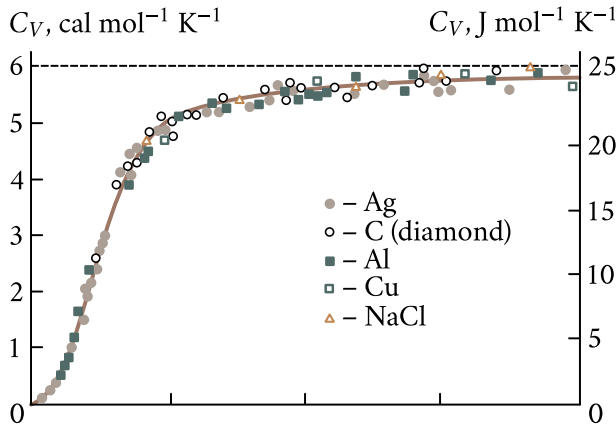


Figure 4.5: Temperature dependence of heat capacity of solids. The solid line is the theoretical Debye curve.

aluminium, copper, and rock salt. The agreement between theory and experiment is quite satisfactory not only from the qualitative but also from the quantitative point of view.

Knowing the temperature dependence of the energy of a lattice, we can easily find at least the qualitative dependence of the concentration of the phonon gas on temperature, that is, the number of phonons n_{ph} excited in a unit of volume of the crystal.

The concentration of the phonon gas in the low temperature range, in which $E_{\text{lattice}} \propto T^4$ and the phonon energy $\hbar\omega \approx k_B T \propto T$, must be proportional to T^3 :

$$n_{\text{ph}} \propto T^3. \quad (\text{low temperature range, } T \ll \Theta) \quad (4.36)$$

In the high temperature range, where $E_{\text{lattice}} \propto T$ and the phonon energy attains the maximum value of $\hbar\omega_D \approx k_B T$ independent of T , the concentration of the phonon gas should be proportional to T :

$$n_{\text{ph}} \propto T. \quad (\text{high temperature range, } T > \Theta) \quad (4.37)$$

To be exact, to calculate the concentration of the phonon gas one must know the average energy of the phonons $\bar{\varepsilon}_{\text{ph}}$ both in the low and the high temperature ranges since the lattice energy is equal to the product of the average phonon energy and their concentration. The calculation of $\bar{\varepsilon}_{\text{ph}}$ yields

$$\bar{\varepsilon}_{\text{ph}} = \frac{\pi^2 k_B T}{5} \quad (4.38)$$

for the low temperature range, and

$$\bar{\varepsilon}_{\text{ph}} = \frac{2k_B \Theta}{3} \quad (4.39)$$

for the high temperature range.

This justifies the temperature dependence of n_{ph} expressed by formulae (4.36) and (4.37) and obtained by qualitative methods.

§ 34. Heat capacity of electron gas

The metals, in addition to ions, which constitute the lattice and vibrate around their equilibrium positions, contain also free electrons, the number of which per unit volume is approximately the same as that of the ions. For this reason, the specific heat of a metal should be the sum of the heats capacity of the lattice $C_{lattice}$ calculated in the previous paragraph and of the electron gas C_e :

$$C_V = C_{lattice} + C_e.$$

If the electron gas was a normal classical (nondegenerate) gas, every electron would have an average energy $3k_B T/2$ and the energy of the electron gas per mole of the metal would be

$$E_e^{(cl)} = \frac{3}{2} N_A k_B T = \frac{3}{2} RT$$

and its heat capacity would be

$$C_e^{(cl)} = \frac{3}{2} N_A k_B = \frac{3}{2} R. \quad (4.40)$$

The total heat capacity of the metal in the high temperature range would in this case be

$$C_V = C_{lattice} + C_e = \frac{9}{2} R \approx 37 \text{ J mol}^{-1} \text{ K}^{-1}.$$

Actually, the heat capacity of metals, as well as that of dielectrics, in the high temperature range, where the Dulong and Petit law is valid, is $C_V \approx 25 \text{ J mol}^{-1} \text{ K}^{-1}$, a proof that the contribution of the electron gas is negligible.

This situation incomprehensible from the point of view of classical physics found its natural explanation in quantum theory.

Indeed, as was demonstrated in Chapter 3, the electron gas in metals is a degenerate gas described by the Fermi-Dirac quantum statistics. As the temperature is raised, not all the electrons are thermally excited; only a negligible fraction of them, ΔN , occupying states close to the Fermi level (see Figure 3.6) are thermally excited. The number of such electrons is approximately expressed by the relation (3.43):

$$\Delta N \approx N \frac{k_B T}{2E_F}$$

where E_F is the Fermi energy. For copper at $T \approx 300 \text{ K}$ and $E_F \approx 7 \text{ eV}$, we have $\Delta N/N \approx 0.002$, that is, less than one percent.

Every thermally excited electron absorbs an energy of the order of $k_B T$ just as a particle of a normal gas does. The energy absorbed by the electron gas as a whole is the product of $k_B T$ and the number of thermally excited electrons ΔN :

$$E_e \approx k_B T \Delta N \approx N k_B T \frac{k_B T}{2E_F}. \quad (4.41)$$

The heat capacity of the electron gas is

$$C_e = \frac{dE_e}{dT} \approx N k_B \frac{k_B T}{E_F}. \quad (4.42)$$

A more accurate calculation yields the following expression:

$$C_e \approx \pi^2 N k_B \frac{k_B T}{2E_F}. \quad (4.43)$$

Comparing Eq. (4.40) with (4.43), we obtain

$$\frac{C_e}{C_e^{(cl)}} \approx \pi \frac{k_B T}{E_F}. \quad (4.44)$$

It follows from Eq. (4.44) that the ratio of the heat capacity of a degenerate electron gas to that of a nondegenerate monatomic gas is approximately equal to the ratio of $k_B T$ to E_F . At normal temperatures, the ratio $\pi k_B T / E_F \lesssim 1\%$. Therefore,

$$C_e \lesssim 0.01 C_e^{(cl)}. \quad (4.44')$$

Hence, because of the degeneracy of the electron gas in metals even in the high temperature range only a small portion of the free electrons (usually less than one percent) is thermally excited; the rest do not absorb heat. This is why the heat capacity of the electron gas is negligible as compared to that of the lattice, and the heat capacity of a metal as a whole is practically equal to the latter.

The situation is different in the low temperature range close to absolute zero. Here, the heat capacity decreases in proportion to T^3 , with the fall in temperature and close to absolute zero may prove to be so small that the contribution of the heat capacity of the electron gas, C_e , which decreases much more slowly than C_{lattice} ($C_e \propto T$), may become predominant. Figure 4.6 shows the temperature dependence of the lattice and electron components of specific heat of an alloy (20% vanadium and 80% chromium) whose Debye temperature is $\Theta = 500$ K. It may be seen from Figure 4.6, that close to absolute zero the heat capacity of the electron gas is much greater than that of the lattice ($C_{\text{lattice}} < C_e$), the sign of the inequality remaining the same up to $T \approx 8.5$ K. At $T > 8.5$ K, the sign is reversed, the inequality becoming stronger with the rise in T . Already at $T \approx 25$ K, the heat capacity of the alloy is mainly due to that of its lattice (at $T = 25$ K the heat capacity is $C_{\text{lattice}} \approx 10C_e$).

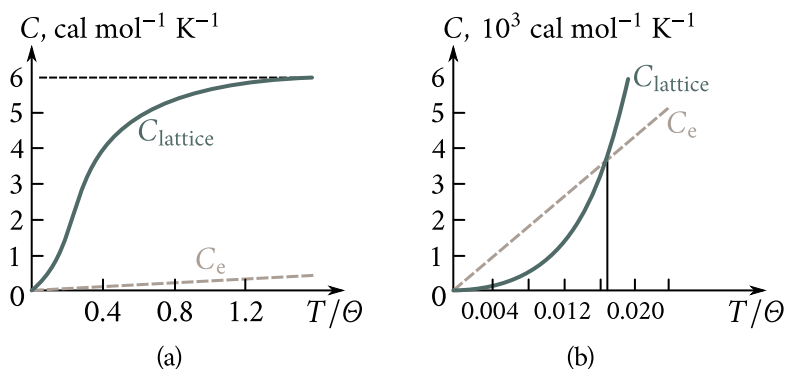


Figure 4.6: Temperature dependence of lattice and of an alloy consisting of 20% vanadium and 80% chromium.

§ 35. Thermal expansion of solids

To explain the elastic properties of solids, in Chapter 2 we have introduced the harmonic approximation according to which the elastic force acting on a particle displaced from its equilibrium position is proportional to the displacement [see Eq. (2.3)] and its potential energy is proportional to the square of the displacement [see Eq. (2.2)]; this fact, is represented by a parabola (the dotted line in Figure 2.1).

The immediate result of the harmonic approximation was the Hooke's law, which describes the elastic deformation of solids. The same approximation was used in Chapter 3 as a basis for calculating the thermal vibrations of a lattice and constructing the theory of heat capacity of a lattice, which is in fair agreement with experiment.

However, the harmonic approximation was unable to explain such well known phenomena as, for instance, the thermal expansion of solids, their heat conductivity, etc.

Indeed, let us turn to the dependence of the potential energy of interaction of the particles of a solid on the distance between them (Figure 4.7). At absolute zero, the particles occupy positions r_0 corresponding to the minimum interaction energy U_0 (at the bottom of the potential trough (well) "abc"). Those distances determine the dimensions of the body at absolute zero. As the temperature rises the particles begin to vibrate around their equilibrium positions 0. For the sake of simplicity, let us assume that particle 1 is fixed and only particle 2 is vibrating. The kinetic energy of the vibrating particle is at its maximum E_k when the particle passes its equilibrium position 0. In Figure 4.7 the energy E_k is measured upwards from the bottom of the potential trough. When particle 2 moves to the left, its kinetic

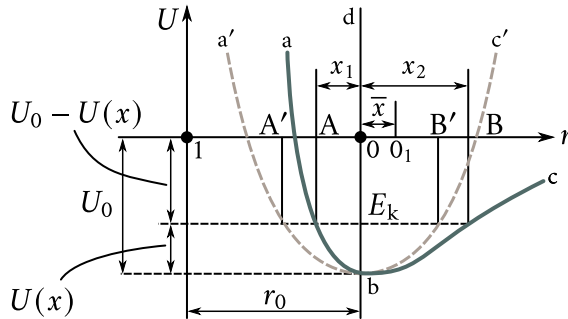


Figure 4.7: The origin of thermal expansion of solids (explanation in text).

energy is used to overcome the repulsive forces acting from particle 1 and is transformed into the potential energy of the particles' interaction. The displacement to the left stops when all the kinetic energy E_k is transformed into the potential energy. In the extreme left position of particle 2 displaced by the distance x_1 , the potential energy's increment is $U(x_1) = E_k$ and its value $-[U_0 - U(x_1)]$. When particle 2 moves to the right, its kinetic energy is spent to overcome the forces of attraction to particle 1 and is, as in the previous case, transformed into the potential energy of the particles' interaction. At point B displaced from the equilibrium position by the distance x_2 , the entire kinetic energy E_k is transformed into the potential energy, the latter increasing by $U(x_2) = E_k$ to become $-[U_0 - U(x_2)]$.

If the vibrations of particle 2 were purely harmonic, the force $f(x)$ caused by its displacement from the equilibrium position by a distance x would be strictly proportional to this displacement and directed towards the equilibrium position:

$$f = -\beta x. \quad (4.45)$$

The change in the particle's potential energy $U(x)$ would, in this case, be described by the parabola $a'bc'$ (Figure 4.7) whose equation is

$$U(x) = \frac{\beta x^2}{2}. \quad (4.46)$$

This parabola is symmetric about the straight line bd passing parallel to the axis of ordinates at a distance of r_0 from it. Therefore, the displacements x_1 and x_2 would be equal in magnitude and the centre of AB would coincide with the equilibrium position 0. In this case, heating a body would not bring about its expansion, for a rise in temperature would result only in an increase in the particles' amplitude of vibrations, the average distances between them remaining unchanged.

Actually, the potential curve abc is, as may be seen from Figure 4.7, not symmetric about the straight line bd , its left branch ba rising much more steeply than the right branch bc . This means that the vibrations of the particles in a solid are

anharmonic (not harmonic). To account for the asymmetry of the potential curve an additional term $-gx^3/3$ expressing this asymmetry (g is a proportionality factor) should be introduced. Then, Eqs. (4.45) and (4.46) will assume the following form:

$$U(x) = \frac{\beta x^2}{2} - \frac{gx^3}{3}, \quad (4.45')$$

$$f(x) = -\frac{\partial U}{\partial x} = -\beta x + gx^2. \quad (4.46')$$

When the particle 2 is displaced to the right ($x > 0$), the term $gx^3/3$ is subtracted from $\beta x^2/2$ and the slope of the branch be is less than that of the branch be' ; when the displacement is to the left ($x < 0$), the term $gx^3/3$ is added to $\beta x^2/2$ and the slope of ba is greater than that of ba' .

Because of the asymmetric nature of the potential curve, the right and left displacements of particle 2 turn out to be different, the former being greater than the latter (Figure 4.7). As a result, the central position of particle 2 (point 0₁) no longer coincides with its equilibrium position 0 but is displaced to the right. This corresponds to an increase in the average distance between the particles by x .

Hence, heating a body should result in an increase in the average distances between particles and the body should expand. The cause of this is the anharmonic nature of the vibrations of particles making up the solid due to the asymmetry of the dependence of the particles' interaction energy on the distance between them.

Let us estimate the value of the thermal expansion coefficient.

The average value of the force caused by the displacement of particle 2 from its equilibrium position is

$$\bar{f} = -\beta \bar{x} + g \bar{x}^2.$$

When the particle vibrates freely, $\bar{f} = 0$; therefore, $g \bar{x}^2 = \beta \bar{x}$. Hence, we obtain

$$\bar{x} = \frac{g \bar{x}^2}{\beta}. \quad (4.47)$$

The expression for the potential energy of a vibrating particle correct to the second order of magnitude is (4.45) and its average value is $\bar{U}(x) \approx \beta \bar{x}^2/2$. Hence,

$$\bar{x}^2 \approx \frac{2\bar{U}(x)}{\beta}.$$

Substituting this into Eq. (4.47), we obtain

$$\bar{x} = \frac{2g\bar{U}(x)}{\beta^2}.$$

In addition to potential energy $U(x)$, a vibrating particle has kinetic energy E_k such that $\bar{U}(x) = \bar{E}_k$. The total average energy of the particle is $\bar{E} = \bar{U}(x) + \bar{E}_k =$

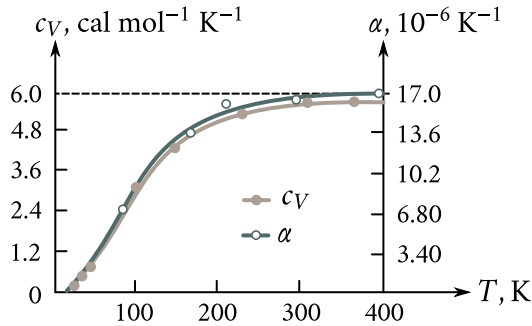


Figure 4.8: Temperature dependence of linear expansion coefficient α and of heat capacity c_V of copper.

$2\bar{U}(x)$. This fact makes it possible to rewrite the expression for x in the following form:

$$\bar{x} = \frac{g\bar{E}}{\beta^2}.$$

The relative linear expansion, that is, the ratio of the variation of the average distance between the particles, x , to the equilibrium distance between them, r_0 , is equal to

$$\frac{\bar{x}}{r_0} = \frac{g}{\beta^2 r_0} \bar{E}$$

and the *linear expansion coefficient* is

$$\alpha = \frac{1}{r_0} \frac{d\bar{x}}{dT} = \frac{g}{\beta^2 r_0} \frac{d\bar{E}}{dT} = \chi c_V \quad (4.48)$$

where

$$\chi = \frac{g}{\beta^2 r_0} \quad (4.49)$$

and c_V is heat capacity per particle.

Thus, the thermal expansion coefficient proves to be proportional to temperature. Figure 4.8 shows the temperature dependence of c_V and α . It can be easily seen that both are interrelated.

In the high temperature range the energy of a particle engaged in linear vibrations is $k_B T$ and its heat capacity is $c_V = k_B$. Therefore, the thermal expansion coefficient of a linear atomic chain will be

$$\alpha = \chi c_V = \frac{g k_B}{\beta^2 r_0}.$$

Substituting the values for g , k_B , β , and r_0 for various solids, we obtain a value of the order of 10^{-4} to 10^{-5} for α , which is in fair agreement with experiment.

Experiment also supports the conclusion that in the high temperature range α is practically independent of temperature (Figure 4.8).

In the low temperature range α behaves in a way similar to that of c_V : it decreases with the fall in temperature and tends to zero as absolute zero is approached.

In conclusion, we would like to remark that for metals a formula similar to (4.48) was first proposed by E. Grüneisen in the form

$$\alpha = \frac{\gamma\kappa}{3V}c_V \quad (4.50)$$

where κ is the metal's compressibility, V the atomic volume, and γ the Grüneisen constant, ranging from 1.5 to 2.5, depending on the metal.

§ 36. Heat conductivity of solids

Heat conductivity of dielectrics (lattice heat conductivity). The second effect caused by the anharmonic nature of atomic vibrations is the *thermal resistance* of solids. There could be no such resistance should the atomic vibrations be strictly harmonic propagating through the lattice in the form of noninteracting elastic waves. In the absence of interaction, the waves would be able to travel without scattering, that is, without meeting any resistance, like light in vacuum.

If we were to set up a temperature difference in such a crystal, the atoms of the hot end vibrating with large amplitudes would transmit their energy to the neighbouring atoms and the front of the heat wave would travel along the crystal with the velocity of sound. Because the wave would meet no resistance there would be a considerable heat flux even for an infinitesimally small temperature difference and the heat conductivity of the crystal would be infinitely great.

The nature of atomic vibrations in real crystals at temperatures not too low is anharmonic, as indicated by the second term in Eq. (4.45'). The anharmonicity destroys the independence of the normal modes of the lattice and causes them to interact, exchanging energy and changing the direction of their propagation (through mutual scattering). It is just those processes of interaction between the elastic waves that make possible the transfer of energy from the modes of one frequency to the modes of another and the establishment of thermal equilibrium in the crystal.

The process of mutual scattering of normal modes is conveniently described in terms of phonons, the thermally excited crystal being regarded as a box containing phonons. In the harmonic approximation, in which the normal modes are presumed to be independent, the phonons make up an ideal gas (a gas of noninteracting phonons). The transition to the anharmonic modes is equivalent to the introduction of an interaction between phonons, which may result in the splitting

of a phonon into two or more phonons and in the formation of a phonon from two other phonons. Such processes are termed *phonon-phonon scattering*. Their probability, as is the case of all scattering processes, is characterized by the *effective scattering cross-section* σ_{ph} . Should the phonon be, from the point of view of the scattering processes, represented by a sphere of the radius r_{ph} then, $\sigma_{\text{ph}} = \pi r_{\text{ph}}^2$. The phonon-phonon scattering may take place only if the phonons approach to within a distance at which their effective cross sections begin to overlap. Since the scattering is due to the anharmonicity of the atomic vibrations, numerically described by the coefficient of anharmonicity g , it would be natural to assume that the phonon effective cross section radius is proportional to g and $\sigma_{\text{ph}} \propto g^2$.

Knowing the effective scattering cross section σ_{ph} , we can calculate the mean free path λ_{ph} of the phonons, that is, the average distance the phonons travel between two consecutive scattering acts. Calculations show that

$$\lambda_{\text{ph}} = \frac{1}{n_{\text{ph}}\sigma_{\text{ph}}} \propto \frac{1}{n_{\text{ph}}g^2} \quad (4.51)$$

where n_{ph} is the phonon concentration.

In the kinetic theory of gases it is proved that for gases the *heat conductivity* is

$$\mathcal{K} = \frac{\lambda \nu C_V}{3} \quad (4.52)$$

where λ is the mean free path of the molecules, ν their thermal velocity, and C_V the heat capacity of the gas.

Let us apply this formula to the phonon gas substituting for C_V the specific heat of the crystal (the phonon gas), for $\lambda = \lambda_{\text{ph}}$ the mean free path of the phonons, and for ν the velocity of sound (the phonon velocity). We obtain the following expression for the lattice heat conductivity:

$$\mathcal{K}_{\text{lattice}} = \frac{\nu \lambda_{\text{ph}} C_V}{3}. \quad (4.53)$$

Substituting λ_{ph} into (4.53) from (4.51), we obtain

$$\mathcal{K}_{\text{lattice}} \propto \frac{\nu C_V}{n_{\text{ph}} g^2}. \quad (4.54)$$

In the high temperature range, in accordance with (4.37), $n_{\text{ph}} \propto T$; hence,

$$\mathcal{K}_{\text{lattice}} \propto \frac{\nu C_V}{T g^2}. \quad (4.55)$$

Since C_V in this range is practically independent of T , the lattice thermal conductivity should be inversely proportional to the absolute temperature, which is in qualitative agreement with experiment. Equation (4.55) also includes the anharmonicity factor g and the sound velocity ν , which depend substantially on the rigidity of the bonds between the particles of the solid. Bonds of lesser rigidity cor-

respond to lower ν 's and to greater g 's, since the weakening of the bonds makes for greater thermal vibration amplitudes (for a specified temperature) and for greater anharmonicity. Both those factors should, according to (4.55), bring about a reduction in $\mathcal{K}_{\text{lattice}}$. This conclusion is supported by experiment. Table 4.2 presents the values of sublimation heat Q_s , which is a measure of bonding energy, and of the lattice heat conductivity lattice for some covalent crystals with the diamond lattice: diamond, silicon, and germanium.

We see that the decrease in the bond energy from the value of diamond to that of silicon and, finally, germanium is accompanied by a noticeable decrease in the lattice heat conductivity. A more detailed analysis shows that $\mathcal{K}_{\text{lattice}}$ is also strongly dependent of the mass M of the particles, being less for greater M 's. This, to a large extent, accounts for the fact that the lattice heat conductivity of the lighter elements occupying the upper part of the Mendeleev periodic table (B, C, Si) is of the order of tens or even hundreds of watts per metre per kelvin, the corresponding values for the elements of the middle part of the Mendeleev table being several watts per metre per kelvin, and that for the heavier elements occupying the lower part of the periodic table even to several tenths of a watt per metre per kelvin.

A striking feature is that the lattice heat conductivity of crystals with light particles and rigid bonds may be very high. Thus, at room temperature $\mathcal{K}_{\text{lattice}}$ of diamond is greater than the total heat conductivity of the best heat conductive metal, silver: $\mathcal{K}_{\text{Ag}} = 407 \text{ W m}^{-1} \text{ K}^{-1}$.

At temperatures below the Debye temperature there is a sharp drop in phonon concentration with a fall in temperature leading to a sharp increase in their mean free path, so that at $T \ll \Theta/20$ it becomes comparable with the dimensions of the crystal. Since the crystal surface usually is a poor reflector of phonons, any further decrease in temperature does not lead to an increase in λ_{ph} , for the latter is determined by the crystal's dimensions only. The temperature dependence of the lattice heat conductivity within this range of temperatures is determined by the temperature dependence of the heat capacity C_V . Since $C_V \propto T^3$ in the low

Table 4.2

Substance	$Q_s, (10^5 \text{ J mol}^{-1})$	$\mathcal{K}_{\text{lattice}}, (\text{W m}^{-1} \text{ K}^{-1})$
Diamond	71.23	550.0
Silicon	46.09	137.0
Germanium	37.00	54.00

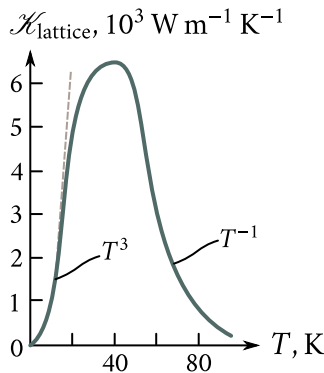


Figure 4.9: Temperature dependence of heat conductivity of synthetic sapphire.

temperature range, $\mathcal{K}_{\text{lattice}}$ too should be proportional to T^3 :

$$\mathcal{K}_{\text{lattice}} \propto T^3. \quad (4.56)$$

This result is also in qualitative agreement with experiment. Figure 4.9 shows the temperature dependence of thermal conductivity of synthetic sapphire. In the low temperature range $\mathcal{K}_{\text{lattice}}$ is indeed approximately proportional to T^3 .

As the temperature rises so does the concentration of phonons n_{ph} , and this should *per se* cause $\mathcal{K}_{\text{lattice}}$ to rise. However, an increase in n_{ph} is accompanied by an increase in the phonon-phonon scattering and a consequent decrease in the mean free path of phonons λ_{ph} , which should result in a decrease in $\mathcal{K}_{\text{lattice}}$. For low n_{ph} , the first factor should be the predominant one and $\mathcal{K}_{\text{lattice}}$ should rise with T . However, starting with a definite concentration n_{ph} , the second factor should assume primary importance and $\mathcal{K}_{\text{lattice}}$ after passing through a maximum should fall with the rise in T . This decrease in the high temperature range is approximately of the $1/T$ type.

Amorphous dielectrics in which the size of regions with a regular structure is of the order of interatomic distances should present a similar picture. Phonon scattering on the boundaries of such regions should be the dominant factor at all T 's and, therefore, λ_{ph} should be independent of T . Because of that the heat conductivity of such dielectrics should be proportional to T^3 in the low temperature range and independent of T in the high temperature range. This is just what is observed in experiment.

However, at present the theory is unable to predict not only the exact values of lattice heat conductivity, $\mathcal{K}_{\text{lattice}}$, but even its order of magnitude.

Heat conductivity of metals. In metals, in contrast to dielectrics, heat is transported not only by phonons but by electrons as well. Therefore, generally the heat conductivity of metals is the sum of the lattice heat conductivity $\mathcal{K}_{\text{lattice}}$

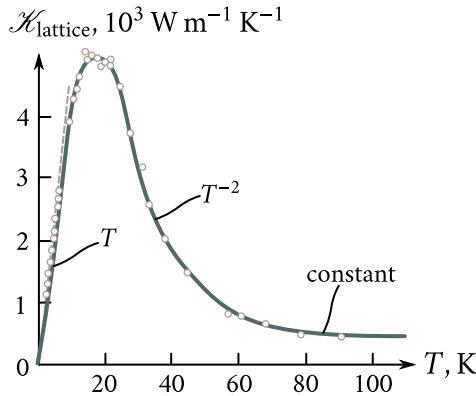


Figure 4.10: Temperature dependence of heat conductivity of copper.

(conductivity due to phonons) and the heat conductivity \mathcal{K}_e of the free electrons:

$$\mathcal{K} = \mathcal{K}_{\text{lattice}} + \mathcal{K}_e.$$

The heat conductivity of the electron gas, can be calculated with the aid of Eq. (4.52). Substituting into this formula the heat capacity of the electron gas, C_e , the electron velocity, v_F , and their mean free path, λ_e , we obtain

$$\mathcal{K}_e = \frac{1}{3} C_e v_F \lambda_e. \quad (4.57)$$

Substituting C_e from Eq. (4.43) into Eq. (4.57), we have

$$\mathcal{K}_e = \frac{\pi^2}{3} \frac{N k_B^2}{m_n v_F} \lambda_e T. \quad (4.58)$$

Let us make a qualitative estimate of the temperature dependence of heat conductivity of pure metals.

High temperature range. Practically, of all the quantities contained in the right-hand side of Eq. (4.58), only λ_e depends on T . For pure metals at temperatures not too low, λ_e is determined by electron-phonon scattering and, all other conditions being equal, is inversely proportional to phonon concentration: $\lambda_e \propto 1/n_{\text{ph}}$. In the high temperature range, according to Eq. (4.37), $n_{\text{ph}} \propto T$. Substituting this into Eq. (4.58), we obtain

$$\mathcal{K}_e = \text{constant}. \quad (4.59)$$

Hence, the heat conductivity of pure metals in the high temperature range should be independent of temperature. This is an experimental fact. Figure 4.10 shows the experimental curve depicting the temperature dependence \mathcal{K} for copper. It follows that above 80–100 K the heat conductivity of copper is practically independent of temperature.

Low temperature range. The phonon concentration in this range is $n_{\text{ph}} \propto T$, therefore, $\lambda_e \propto 1/T^3$. Substituting into Eq. (4.58), we obtain

$$\mathcal{K}_e \propto T^{-2}. \quad (4.60)$$

Consequently, in the low temperature range where the Debye T^3 law is true, the heat conductivity of metals should be inversely proportional to the square of the absolute temperature. This conclusion too is in general supported by experiment (Figure 4.10).

Very low temperature range. Close to absolute zero the phonon concentration in a metal becomes so small that the main part in electron scattering processes is taken over by the impurity atoms, which are always present in a metal no matter how pure it is. In this case, the mean free path of an electron, $\lambda_e \propto 1/N_i$ (N_i is the concentration of impurity atoms), is no longer dependent on temperature and the heat conductivity of a metal becomes proportional to T :

$$\mathcal{K}_e \propto T \quad (4.61)$$

which is an experimental fact.

Let us estimate the magnitude of \mathcal{K}_e for metals, making use of Eq. (4.57). For typical metals, we have $C_e \approx 0.01C_V \approx 3 \times 10^4 \text{ J m}^{-3} \text{ K}^{-1}$, $v_F \approx 10^6 \text{ m s}^{-1}$, and $\lambda_e \approx 10^{-8} \text{ m}$. Substituting into Eq. (4.57), we obtain $\mathcal{K}_e \approx 10^2 \text{ W m}^{-1} \text{ K}^{-1}$. Thus, \mathcal{K}_e for metals may be as high as hundreds of watts per metre per kelvin. This is substantiated by experiment. Table 4.3 shows the room temperature heat conductivities for some typical metals and for one alloy, constantan, which consists of 60% copper and 40% nickel.

It follows that for pure metals \mathcal{K} can indeed be as high as hundreds of watts per metre per kelvin.

Let us also estimate the contribution of the lattice heat conductivity of a metal, making use of Eqs. (4.53) and (4.57):

$$\frac{\mathcal{K}_{\text{lattice}}}{\mathcal{K}_e} = \frac{C_V v \lambda_{\text{ph}}}{C_e v_F \lambda_e}$$

Table 4.3

Metal	$\mathcal{K}, (\text{W m}^{-1} \text{ K}^{-1})$	Metal	$\mathcal{K}, (\text{W m}^{-1} \text{ K}^{-1})$
Silver	403	Aluminium	210
Copper	384	Nickel	60.0
Gold	296	Constantan	23.0

v being the phonon (sound) velocity. For pure metals, we have $C_e/C_V \approx 0.01$, $v = 5 \times 10^3 \text{ m s}^{-1}$, $\lambda_{\text{ph}} \approx 10^{-9} \text{ metre}$, $v_F \approx 10^6 \text{ m s}^{-1}$, and $\lambda_e \approx 10^{-8} \text{ m}$. Hence, $\mathcal{K}_{\text{lattice}}/\mathcal{K}_e \approx 5 \times 10^{-2}$.

It follows then that the heat conductivity of typical metals is almost entirely due to the heat conductivity of their electron gas, the contribution of lattice heat conductivity being a few percent. This picture may, however, totally change when we go over to metallic alloys, in which impurity scattering is the principal electron scattering mechanism. The electron mean free path for which such scattering is responsible, λ_e , is inversely proportional to the impurity concentration N_i ($\lambda_e \propto 1/N_i$) and for high N_i 's may become comparable with the phonon mean free path ($\lambda_{\text{ph}} \propto \lambda_e$). Naturally, in such a case the electron contribution to heat conductivity may become of the same order of magnitude as that of the phonon contribution: $\mathcal{K}_e \approx \mathcal{K}_{\text{lattice}}$. This too is supported by experiment. Table 4.3 gives the heat conductivity of constantan. It is much less than that of nickel or copper. This proves the fact that electron scattering in constantan is mainly due to lattice defects caused by impurity atoms. We also note that \mathcal{K}_e and $\mathcal{K}_{\text{lattice}}$ measured by R. Berman in constantan proved to be of the same order of magnitude.

Chapter 5

The Band Theory of Solids

The theory of free electrons was the first successful attempt to explain the electric and magnetic properties of solids (primarily of metals). It was based on the assumption that metals contain free electrons capable of moving around the metal like gas molecules in a vessel. The theory of free electrons was successful in explaining such phenomena as the electric and the heat conductivities, thermionic emission, thermoelectric and galvanomagnetic effects, etc. However, this theory proved incapable of dealing with such properties of solids as are determined by their internal structure. It could not even explain why some bodies are conductors and some insulators.

The next stage in the progress of the electron theory has been the band theory of solids, which is outlined in this chapter.

§ 37. Electron energy levels of a free atom

The state of an electron in an atom is determined by four quantum numbers: the principal n , the orbital l , the magnetic m_l , and the spin σ numbers. In a hydrogen atom the *principal quantum number*, n , describes the steady-state energy of the electron:

$$E(n) = -\frac{R}{n^2} \quad (5.1)$$

where $R = 13.6 \text{ eV}$ is a universal constant, called a *rydberg*—or the *Rydberg constant*.

The *orbital quantum number*, l , describes the orbital angular momentum of the electron, \mathbf{p}_l :

$$p_l = \hbar (l(l+1))^{1/2} \quad (5.2)$$

($\hbar = h/2\pi$, with h the Planck constant). The quantum number l may assume only

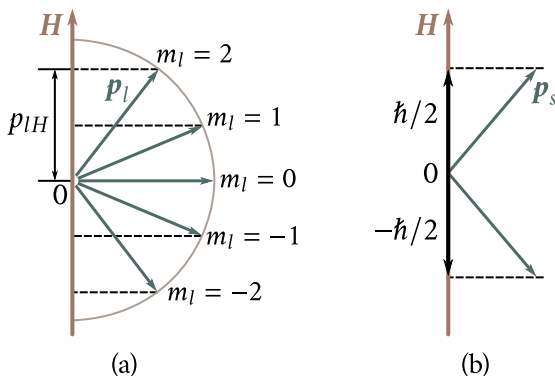


Figure 5.1: Orientations of orbital (a) and spin (b) angular momenta with respect to H .

the following integral values:

$$l = 0, 1, 2, \dots, (n - 1)$$

n value in all.

The *magnetic quantum number*, m_l , describes the orientation of the orbital angular momentum with respect to some specified direction H [Figure 5.1(a)]: the orientation of \mathbf{p}_l with respect to H may only be such that its projection onto this direction is a multiple of \hbar :

$$p_{lH} = m_l \hbar. \quad (5.3)$$

The number m_l may assume the following set of integral values:

$$m_l = -l, -(l - 1), \dots, 0, 1, 2, \dots, l. \quad (5.3')$$

$2l + 1$ values in all.

Lastly, the *spin quantum number*, σ , describes the orientation of the intrinsic angular momentum (the spin \mathbf{p}_s) of the electron with respect to specified direction H [Figure 5.1(b)]: the vector \mathbf{p}_s may only be oriented with respect to H so that its projection onto H is equal to

$$p_{sH} = \sigma \hbar, \quad (5.4)$$

where only the values $+1/2$ and $-1/2$ being allowed for σ .

The states with orbital quantum number $l = 0$ (the values of other quantum numbers being irrelevant) are termed *s states*; those with $l = 1$, are termed *p states*; with $l = 2$, *d states*; $l = 0$, *f states*; etc. Electrons in those states are termed s-, p-, d-, f-, etc. electrons.

In contrast to the hydrogen atom, the energy of an electron in many electron atoms depends not only on n but on l as well: $E = E(n, l)$. Only discrete values of n and l being allowed, the energy spectrum of electrons in atoms may assume

only discrete values too; the spectrum consists of a set of allowed levels $E = E(n, l)$ separated by forbidden energy intervals. Table 5.1 shows a diagram (not to scale) of the first three groups of such levels.

All the s levels are *nondegenerate*. This means that everyone of them corresponds to a single electron state in the atom. In compliance with the Pauli exclusion principle there may be two electrons with opposite spins in such a state.

The p levels are three-fold degenerate: there is not one but three states with different magnetic quantum numbers m_l to correspond to each of them. For $l = 1$ those values are $m_l = -1, 0, +1$. Figure 5.2 shows the shape of electron clouds corresponding to those states. Since there may be two electrons per state, the total number of electrons in the p state is six.

The degeneracy of the d levels is five-fold, since the allowed values of the magnetic quantum number for $l = 2$ are $m_l = -2, -1, 0, +1, +2$. This level can accommodate 10 electrons. Generally, a level with the orbital quantum number l is a $(2l + 1)$ -fold degenerate one and can accommodate $2(2l + 1)$ electrons.

Table 5.1

Atomic energy levels and their notation	$g = 2l + 1$	Total number of electrons on a level: $n = 2(2l + 1)$	Splitting of levels into $g = 2l + 1$ sublevels when degeneracy is lifted
$E(3, 2), 2p$	5	10	<div><div>— 2</div><div>— 1</div><div>3d — 0</div><div>— -1</div><div>— -2</div></div>
$E(3, 1), 3p$	3	6	<div><div>— 1</div><div>3p — 0</div><div>— -1</div></div>
$E(3, 0), 3s$	1	2	<div>3s — 0</div>
$E(2, 1), 2p$	3	6	<div><div>— 1</div><div>2p — 0</div><div>— -1</div></div>
$E(2, 0), 2s$	1	2	<div>2s — 0</div>
$E(1, 0), 1s$	1	2	<div>1s — 0</div>

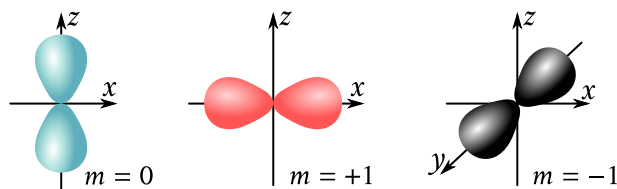


Figure 5.2: Electron clouds of the 2p state.

When a free atom is placed in a strong external field, the degeneracy vanishes and every level splits into $(2l + 1)$ closely spaced sublevels, as shown in the last column of Table 5.1.

The effect of an external field on different atomic levels is not the same. The splitting of the levels of inner electrons, whose interaction with the nucleus is strong, is so small that it may be neglected. As the shell radius is increased the energy of interaction of the respective electrons with the nucleus becomes smaller and the effect of the external field becomes more noticeable. The effect of an external field is most pronounced for the energy levels of outer electrons, whose bonds with the nucleus are relatively weak.

§ 38. Collectivization of electrons in a crystal

The interatomic distances in solids are so small that every atom finds itself in a strong field of the neighbouring atoms. To gain insight into the effect this field exercises on the energy levels, consider the following idealized example.

Arrange N sodium atoms in the pattern of a three-dimensional lattice having the shape of a sodium crystal but with interatomic distances so great that the interaction between the atoms can be neglected. In this case, one can legitimately assume the energy states in every atom to be the same as in an individual sodium atom. Figure 5.3(a) shows the energy diagram of two such atoms. Each of them has the appearance of a spindle-shaped potential trough inside of which the levels 1s, 2s, 2p, 3s, ..., are shown. The 1s, 2s and 2p levels in a sodium atom are fully occupied, the 3s is occupied to one half, and the levels above 3s are empty.

As shown in Figure 5.3(a) the individual atoms are separated by potential barriers of the width $r \gg a$, where a is the lattice constant. The height of the potential barrier U is not the same for electrons occupying different levels. It is equal to the height measured from those levels to the zero level ∞ . The potential barrier prevents the electrons from moving freely from one atom to another. Calculations show that for $r \approx 30 \text{ \AA}$ the average transition rate of a 3s electron from one atom to another is once in every 10^{20} years.

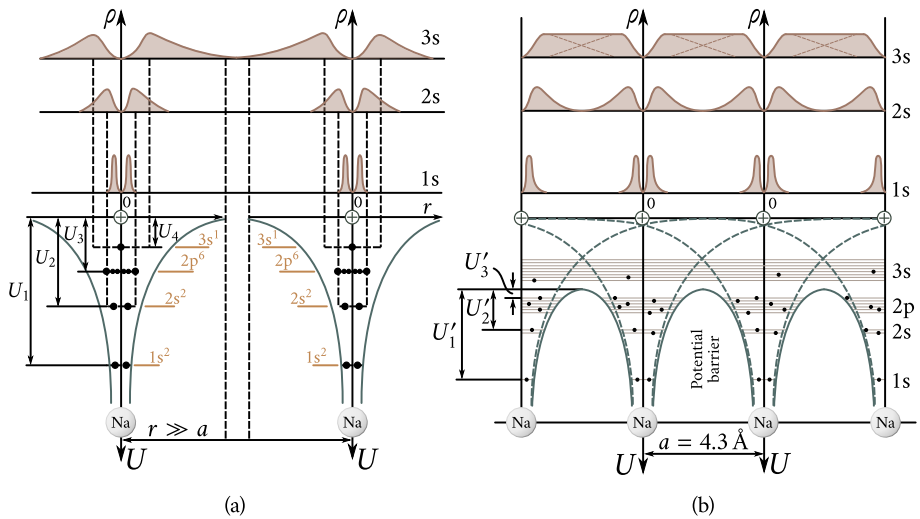


Figure 5.3: Variation of electronic states in approaching atoms: (a)—energy diagram of sodium atoms placed at a distance much greater than the sodium lattice parameter; (b)—energy diagram of sodium atoms brought together to a distance of the order of the lattice parameter.

The upper part of Figure 5.3(a) shows a qualitative picture of the space distribution of the density $\rho = 4\pi\psi\psi^*$ of the probability of detecting electrons at a distance r from the nucleus. The maxima of those curves correspond approximately to the radii of Bohr orbits of such electrons.

Now, make the lattice contract uniformly so that its symmetry remains unaffected. As the atoms are brought closer the interaction between them increases and for an interatomic distance equal to the lattice constant a it turns into one characteristic of the crystal. Figure 5.3(b) depicts that situation. We see that the walls of the potential troughs separating neighbouring atoms (they are shown in the figure by dotted lines) partly overlap to create new walls shown by solid lines that are lower than the zero level 00. Namely, the effect of the reduction of the interatomic distances is two-fold: to reduce the height and the width of the barrier, the latter to the value of the order of a . The height of the reduced barrier is U'_1 for $1s$ -, U'_2 for $2s$ -, and U'_3 for $2p$ -electrons, the original $3s$ levels of the sodium atoms lying above it. This fact makes it possible for the valence electrons of this level to move practically unhindered from one atom to another. The nature of the electron clouds of valence electrons shown in the upper part of the figure also points to the same conclusion: their overlapping is so complete that the density of the resulting cloud is practically uniform [Figure 5.3(b)]. This corresponds to their complete

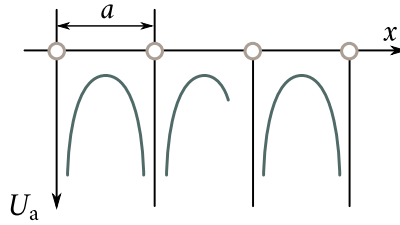


Figure 5.4: Periodic variation of electron potential energy in crystal.

collectivization in the lattice. Such collectivized electrons are usually termed *free electrons* and the totality of them, the *electron gas*.

A drastic reduction in the width and the height of the potential barrier brought about by the decrease in interatomic distances makes it possible for the electrons occupying other atomic levels, besides the valence levels, to move inside the crystal. Their motion takes place by means of tunneling through the barriers that separate neighbouring atoms. The narrower and the lower those barriers are, the more complete is the collectivization of the electrons and the greater is their freedom.

§ 39. Energy spectrum of electrons in a crystal

In the same way as the main goal of the theory of atoms is to describe the electron states in an atom and calculate the allowed energy levels, one of the main problems of the theory of the solid state is to determine the energy spectrum of the electrons in a crystal. One may obtain a qualitative idea about this spectrum using the following approximate method to treat the behaviour of electrons in a crystal.

To approximately describe the motion of an electron in a crystal we, may use the following Schrödinger equation:

$$\nabla^2 \psi + \frac{2m}{\hbar} (E - U) \psi = 0 \quad (5.5)$$

where U is the potential energy, E the total energy, and m the mass of the electron.

For electrons sufficiently strongly bound to the atoms the potential energy may be written in the form

$$U = U_a + \delta U \quad (5.6)$$

where U_a is the electron's energy in an isolated atom. In a crystal it is a periodic function with a period equal to the lattice parameter, since there is a recurrence in the value of energy as the electron moves from one atom to another (Figure 5.4); δU represents a correction that takes account of the effect of neighbouring atoms on this energy.

If one neglects the correction δU in Eq. (5.6), that is, considers only the zero

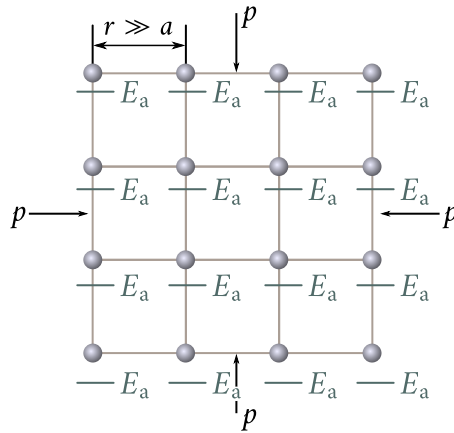


Figure 5.5: Each atomic energy level E_a in a system consisting of N atoms is repeated N times (is N -fold degenerate).

approximation, one should take the wave function ψ_a and the energy $E_a(n, l)$ of the electron in an isolated atom as the wave function and the energy of the electron in a crystal: $\psi = \psi_a$, $E = E_a(n, l)$, where n and l are the principal and orbital quantum numbers, which determine the energy of the electron in an atom.

In this case, the difference between a crystal and an isolated atom is that in an isolated atom a specified energy level $E_a(n, l)$ is unique, but in a crystal consisting of N atoms there are N such levels (Figure 5.5). In other words, every energy level of an isolated atom is N -fold degenerate in a crystal. Such degeneracy is termed *transpositional*.

Now let us estimate the correction δU in the potential energy (5.6). As the isolated atoms are brought together to form a lattice, each atom increasingly feels the field of its neighbours with whom it interacts. As we have already seen, such interaction results in the lifting of degeneracy including transpositional degeneracy. Because of that, each level nondegenerate in an individual atom splits up into N closely spaced sublevels to form an *energy band*.

If every level in an isolated atom was $(2l + 1)$ times degenerate, the corresponding band shall contain $N(2l + 1)$ sublevels. Accordingly, the s level produces the s band consisting of N sublevels and capable of carrying $2N$ electrons; the p level produces the p band consisting of $3N$ sublevels and capable of carrying $6N$ electrons; etc.

The separation between the sublevels in an energy band of an ordinary crystal is very small. A crystal of a volume of one cubic metre contains 10^{28} atoms. For a band of the order of 1 eV wide, the separation between the sublevels is about

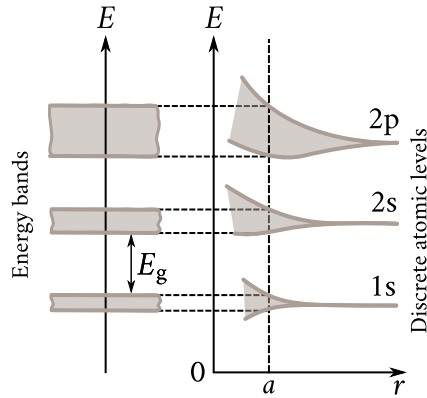


Figure 5.6: Schematic representation of energy band formation in a crystal from discrete atomic levels.

10^{-28} eV. This separation is so negligible that the band may be considered to be practically continuous. However, the fact that the number of levels in a band is finite is very important for the determination of the distribution of electrons over states.

The maximum effect of the lattice field is on the external valence electrons. Because of that, the state of such electrons in a crystal experiences the greatest change and the energy bands formed by their energy levels are the widest. The internal electrons, which are strongly bound to their nuclei, are only slightly perturbed by the presence of neighbouring atoms and accordingly their energy bands in the crystal are almost as narrow as the levels of isolated atoms. Figure 5.6 shows a schematic diagram of energy band formation from discrete atomic levels.

Thus, in a crystal there is an *allowed energy* band to correspond to each energy level of an isolated atom: the 1s band to correspond to the 1s level, the 2p band to the 2p level, etc. The allowed energy bands are separated by *forbidden energy bands* E_g . As the energy of an electron in an atom is increased so is the width of the respective energy band, the width of the forbidden band being reduced.

Figure 5.7 shows the energy band structure of lithium, beryllium, and elements having the diamond lattice (diamond, silicon, germanium). In the lithium crystal [Figure 5.7(a)], the splitting of the 1s level is a narrow one, the splitting of the 2s level being wider so that a sufficiently wide 2s energy band is formed. In the beryllium crystal [Figure 5.7(b)], the 2s and the 2p bands overlap to form a mixed, or the so-called *hybrid*, band. The situation is quite similar in case of the other elements of the main subgroup of Group II of the Mendeleev periodic table.

The pattern of band formation in crystals with the diamond lattice [Figure

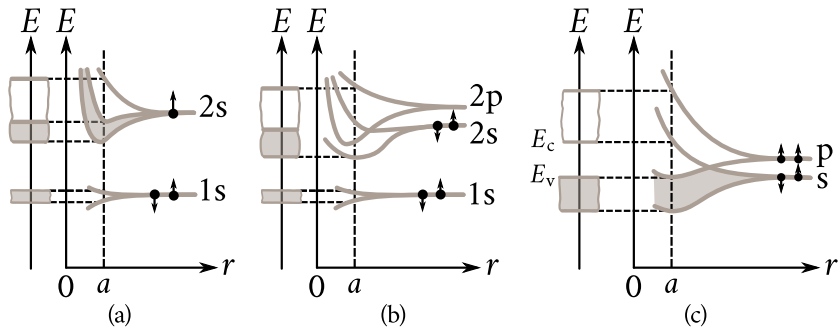


Figure 5.7: Formation of bands from atomic levels: (a)—in lithium crystal (band 2s is only half-filled); (b)—in beryllium crystal (filled 2s band and free 2p band overlap to form a partially filled hybrid band); (c)—in diamond-lattice elements of Group IV (the arrows denote the spin of the electrons).

5.7(c)] is somewhat different. In this case, the bands formed from the s and the p levels overlap and split into two bands, so that each of them contains four states per atom: one s state and three p states. Those bands are separated by a forbidden band. The lower band is termed the *valence band* and the upper the *conduction band*.

§ 40. Dependence of electron energy on the wave vector

It was demonstrated in the preceding section that the pattern of the electron energy spectrum in a crystal is of a band type. Consider now the dependence of the electron energy E on its momentum p inside each band, that is, the shape of the $E(p)$ curves. The momentum dependence of E is termed the *dispersion law*, or *dispersion relation*.

Turn now to the simplest case of a free electron moving along the x axis and described by the following Schrödinger equation:

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar} E\psi = 0 \quad (5.7)$$

where

$$E = \frac{p^2}{2m} \quad (5.8)$$

since a free electron has only kinetic energy.

Formula (5.8) is the *dispersion relation for free electrons*, which expresses the momentum dependence of E . It may be rewritten in the following form. According to the de Broglie formula

$$p = \frac{h}{\lambda} = \frac{\hbar}{(\lambda/2\pi)} = \hbar k \quad (5.9)$$

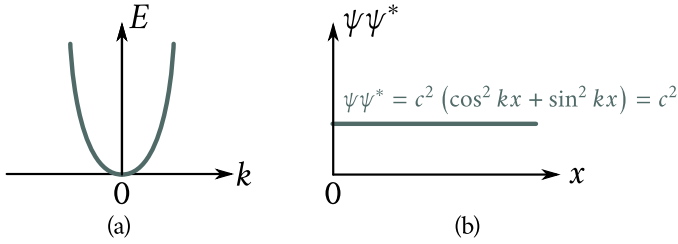


Figure 5.8: Motion of a free electron: (a)—dependence of energy on wave vector (dispersion curve); (b)—square modulus of wave function proportional to the probability of the electron being at point x .

where λ is the wavelength of the electron, and

$$k = \frac{2\pi}{\lambda}. \quad (5.10)$$

The vector \mathbf{k} coinciding in direction with the direction of the electron wave propagation and equal in magnitude to $2\pi/\lambda$ is termed *wave vector of the electron*. Substituting p from Eq. (5.9) into (5.8), we obtain

$$E = \frac{\hbar^2 k^2}{2m}. \quad (5.11)$$

Equation (5.11), which expresses the dependence of the energy of a free electron on its wave vector, is just another form of writing the dispersion relation (the dispersion law) for such electrons.

It follows from Eqs. (5.8) and (5.11) that the dispersion law for free electrons is quadratic and for one-dimensional motion of the electron takes the form of a quadratic parabola shown in Figure 5.8(a).

The solution of Eq. (5.7) is a travelling plane wave

$$\psi = Ce^{ikx} \quad (5.12)$$

where C is the wave's amplitude.

As is well known, the square of the modulus of the wave function is proportional to the probability of detecting the electron in a specified region of space. As may be seen from Eq. (5.12), for a free electron this probability is independent of the electron's position since

$$|\psi|^2 = \psi\psi^* = C^2 \quad (5.13)$$

which means that for a free electron every point in space is equivalent and the probability of detecting it is everywhere the same (Figure 5.8(b)).

The case of an electron moving in a periodic field of a crystal formed by regularly arranged ions is different (Figure 5.9). The probability of detecting it in a specified point of the crystal should be a periodic function in x , since positions

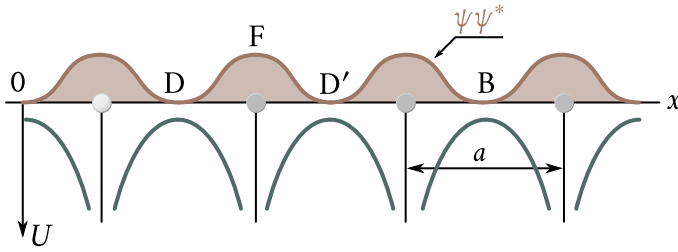


Figure 5.9: Motion of an electron in periodic field. The square modulus of the wave function that describes the probability of the electron being at point x of the horizontal axis is a periodic function of coordinate x , the period being equal to the lattice parameter a .

displaced from one another by a multiple of the lattice constant a (for instance, the positions D, D' and B in Figure 5.9) are equiprobable for the electron. The positions inside a period a (for example, inside DFD') are, however, all different. This means that the amplitude of the wave function $\psi(x)$ of an electron moving in a periodic field does not, as in the case of a free electron, stay constant but changes periodically, or it may be said to be modulated with a period equal to the lattice parameter a . Denote this amplitude $u(x)$. Then, the wave function of the electron moving in a periodic field of a crystal in the direction of the x axis may be expressed in the following form:

$$\psi(x) = u(x)e^{ikx} \quad (5.14)$$

where $u(x + na) = u(x)$, where n is an arbitrary integer. Relation (5.14) is termed the *Bloch function*. The specific form of this function is determined by the potential energy $U(x)$, which enters into the Schrödinger equation (5.5).

There should be a corresponding change in the dispersion relation for electrons moving inside a crystal as compared with that for free electrons. Firstly, as we have already seen, the energy spectrum of such electrons assumes a band pattern: allowed energy bands formed from corresponding atomic levels E_a are separated by forbidden energy bands. Secondly, calculations show the electron energy inside each band to be a function of the wave vector k , which for a one-dimensional crystal (an atomic chain) with the parameter a is of the form

$$E(k) = E_a + C + 2A \cos(ka) \quad (5.15)$$

where E_a is the energy of the atomic level from which the band was formed; C is the displacement of this level due to the effect of the field of neighbouring atoms; A is the so-called *exchange integral*, which takes into account the newly created probability for an electron to move from one atom to another owing to the overlapping of the atomic wave functions [Figure 5.3(b)]. The exchange integral is the greater the greater the overlapping of the wave functions, that is, the greater is the

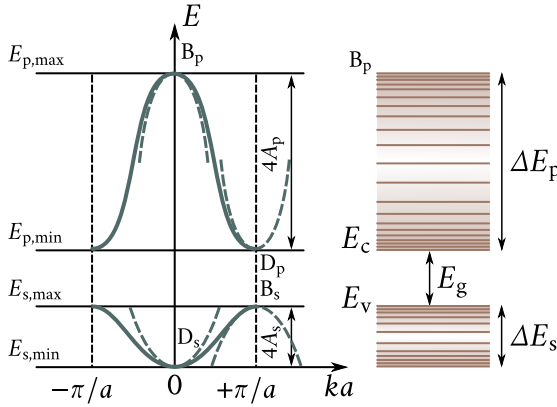


Figure 5.10: Dispersion curves for an electron moving in a periodic field: the lower curve corresponds to the s band, the upper curve to the p band; dotted lines are parabolas expressing the $E(k)$ dependence in the centre and on the boundaries of the Brillouin zone.

exchange rate of the electrons of neighbouring atoms. For s states $A_s < 0$, for p states $A_p > 0$, therefore, it is reasonable to write out relation (5.15) individually for the s and the p bands. For the s bands

$$E_s(k) = E'_s - 2A_s \cos(ka) \quad (5.16)$$

and for the p bands

$$E_p(k) = E'_p - 2A_p \cos(ka) \quad (5.17)$$

where $E'_s = E_{as} + C_s$, $E'_p = E_{ap} + C_p$, and A_s and A_p are the absolute values of the exchange integrals for the respective states.

Figure 5.10 shows dispersion curves $E(k)$ for the s and p bands drawn to satisfy equations (5.16) and (5.17). For the s states, E_s at $k = 0$ assumes its minimum value $E_{s,min} = E'_s - 2A_s$. As k increases, $\cos(ka)$ decreases and the value of $E_s(k)$ rises reaching its maximum $E_{s,max} = E'_s + 2A_s$ at $k = \pi/a$. In the interval of values of k from 0 to $-\pi/a$, $E_s(k)$ changes in a similar fashion. The width of the allowed s band from $E_{s,min}$ to $E_{s,max}$ is

$$\Delta E_s = E_{s,max} - E_{s,min} = 4A_s. \quad (5.18)$$

As may be seen, it is determined by the absolute value of the exchange integral, which, in its turn, depends on the overlapping of wave functions of neighbouring atoms. The shape of the curve $E_s(k)$ is that of an overturned bell.

For the p states $E_{p,min} = E'_p - 2A_p$ corresponds to $k = \pm\pi/a$, and at $k = 0$, $E_{p,max} = E'_p + 2A_p$. The width of the p band

$$\Delta E_p = E_{p,max} - E_{p,min} = 4A_p \quad (5.19)$$

as in the previous case, is determined by the absolute value of the exchange integral A_p . As a rule, the higher the atomic level the greater the overlapping of the corresponding wave functions in the crystal, the greater the value of the exchange integral, and the wider the energy band formed from this level. For this reason, high atomic levels are the origin of wide energy bands separated by narrow forbidden bands (see Figure 5.6).

The intervals of k inside which the electron energy $E(k)$ as a periodic function of k completes its full cycle are termed *Brillouin zones*. For a one-dimensional crystal (an atomic chain) the first Brillouin zone lies between $k = -\pi/a$ and $k = \pi/a$ and is $2\pi/a$ wide (Figure 5.10). In the vicinity of a dispersion curve's extremum, that is, in the vicinity of $k = 0$ and $k = \pm\pi/a$ (the centre and the boundary of the first Brillouin zone), $\cos(ka)$ can be expanded into a power series in ka (k is measured from 0 if the extremum is in the centre of the Brillouin zone and from $\pm\pi/a$ if it is on its boundary) leaving only two terms of the expansion

$$\cos(ka) = 1 - \frac{(ka)^2}{2} + \dots$$

Substituting this into Eqs. (5.16) and (5.17), we obtain $E_s(k) = E_{s,\min} + A_s(ka)^2$ and $E_p(k) = E_{p,\max} - A_p(ka)^2$. The minimum of the dispersion curve $E(k)$ is termed the *bottom of the energy band* and the maximum the *top of the band*. Therefore, we may rewrite the sought for relations in a more general form:

$$E_{\text{bottom}}(k) = E_{\min} + A(ka)^2 \quad (5.20)$$

for the bottom of the band, and

$$E_{\text{top}}(k) = E_{\max} - A'(ka)^2 \quad (5.21)$$

for the top of the band.

Hence, close to the top and the bottom of an energy band the portion of the electron energy that depends on the wave vector is proportional to the square of the wave vector, measured in the way indicated above, and to the exchange integral that determines the width of the band. The parabolas corresponding to equations (5.20) and (5.21) are shown in Figure 5.10 by dotted lines.

The $E(k)$ dependence for real crystals is, as a rule, much more complex than that expressed by Eq. (5.15).

Figure 5.11(a) shows the dispersion curves for the bottom of the conduction band (curve 1) and the top of the valence band (curve 2) of silicon. We see that the bottom of the conduction band, D, of silicon is not in the centre of the Brillouin zone but near its boundary in the direction $[100]$. The valence band is bounded by a curve in the shape of a parabola with its apex at B in the centre of the Brillouin zone. However, despite such complexity of the dispersion curves, the quadratic

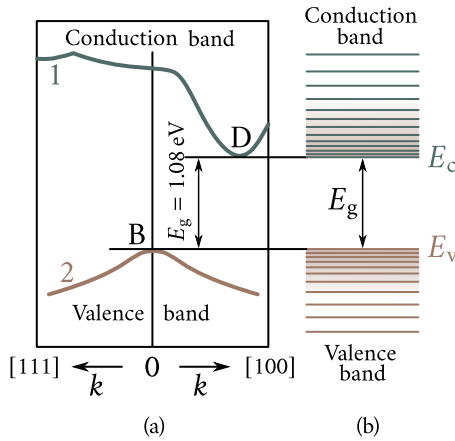


Figure 5.11: Band pattern of silicon: (a)—dispersion curves $E(k)$ bounding the conduction band (curve 1) and the valence band (curve 2); the energy minimum of the conduction band is at point D in the [100] direction, energy maximum of the valence band is in the Brillouin zone centre B; the distance between minimum D and maximum B is the forbidden band width E_g ; (b)—schematic representation of energy band pattern of silicon.

dependence of $E(k)$ expressed by formulae (5.20) and (5.21) remains valid for both band-edges in this case.

The width of the forbidden band, or *energy gap*, is determined by the minimum gap between the valence and the conduction bands; in Figure 5.11(a) this is denoted by E_g .

Frequently, when making a simplified analysis of the energy-band structure of semiconductors, instead of the actual dispersion curves, which bound the valence and the conduction band, use is made of two parallel lines, one drawn tangentially to the bottom of the conduction band and the other to the top of the valence band [Figure 5.11(b)]. The first line is taken to represent the lower boundary (the bottom) of the conduction band and the second the upper boundary (the top) of the valence band. The separation between the lines is equal to the forbidden band width E_g .

§ 41. Effective mass of the electron

The de Broglie formula establishes the following relation between the momentum of a free electron and its wave vector:

$$\mathbf{p} = \hbar \mathbf{k}.$$

The velocity of the electron's translational motion is

$$\mathbf{v} = \frac{\mathbf{p}}{m} = \left(\frac{\hbar}{m} \right) \mathbf{k}. \quad (5.22)$$

Differentiating (5.11) with respect to k , we obtain

$$k = \frac{m}{\hbar^2} \frac{dE}{dk}.$$

Substituting this into (5.9) and (5.22) we obtain

$$p = \hbar k = \frac{m}{\hbar} \frac{dE}{dk}, \quad v = \frac{\hbar}{m} k = \frac{1}{\hbar} \frac{dE}{dk}. \quad (5.23)$$

Expressions for the momentum and for the velocity of translational motion written in this form turn out to be valid not only for free electrons but for electrons moving in a periodic crystal field as well. Momentum \mathbf{p} is in the latter case termed *quasimomentum* of the electron.

Set up an external field \mathcal{E} in the crystal. This field acts on the electron with a force

$$\mathbf{F} = -q\mathcal{E}$$

imparting to it an acceleration

$$j = \frac{dv}{dt} = \frac{1}{\hbar} \frac{d}{dt} \left(\frac{dE}{dk} \right) = \frac{1}{\hbar} \frac{d^2 E}{dk^2} \frac{dk}{dt}.$$

The work performed by the force F during the interval dt is

$$dW = Fv dt = \frac{F}{\hbar} \frac{dE}{dk} dt.$$

This work is spent on increasing the electron's energy by an amount dE , where $dE = (F/\hbar)(dE/dk) dt$. Hence, $F/\hbar = dk/dt$. Substituting into the right-hand side of the expression for j , we obtain

$$j = \frac{F}{\hbar^2} \frac{d^2 E}{dk^2} = -\frac{q\mathcal{E}}{\hbar^2} \frac{d^2 E}{dk^2} = -\frac{q\mathcal{E}}{m_{\text{eff}}}. \quad (5.24)$$

Equation (5.24) establishes the relation between the electron's acceleration j and the external force F with which an external field \mathcal{E} acts on it. Hence, it is an expression of Newton's second law. It follows then that the electron acted upon by an external force moves in a periodic crystal field on the average in the same way as a free electron would move if its mass were

$$m_{\text{eff}} = \hbar^2 \left(\frac{d^2 E}{dk^2} \right)^{-1}. \quad (5.25)$$

The mass m_{eff} is called the *effective mass of the electron*. Having attributed to the electron in a periodic crystal field a mass m_{eff} , we may now regard it as being free and describe its motion in an external field in the same way as we would describe

the motion of an ordinary free electron.

However, the effective mass, which embraces all the details of electron motion in a periodic crystal field, is a very particular quantity. To begin with, it may be positive or negative, many times larger or many times smaller in magnitude than the electron's rest mass m . Let us make a more detailed study of the problem.

For electrons close to the bottom of a band the energy is $E_{\text{bottom}} = E_{\text{min}} + A(ka)^2$, the second derivative with respect to k being $d^2E/dk^2 = 2Aa^2$. Substituting into Eq. (5.25), we obtain the following expression for the effective mass of the electron, which we shall denote m_n :

$$m_n = \frac{\hbar^2}{2Aa^2}. \quad (5.26)$$

Since $A > 0$, we see that $m_n > 0$. Hence, electrons close to the bottom of an energy band have a positive effective mass. For this reason, they behave normally in an external field, accelerating in the direction of the acting force. The difference between such electrons and free electrons is that their mass may be quite different from the rest mass. It may be seen from Eq. (5.26) that the greater A is, that is, the wider the allowed band, the less the effective mass of the electrons occupying states close to the bottom of the band is.

For electrons close to the top of the band the energy is $E_{\text{top}} = E_{\text{max}} - A'(ka)^2$, the second derivative of E with respect to k being $d^2E/dk^2 = -2A'a^2$, and the effective mass, which we denote m'_n , is

$$m'_n = -\frac{\hbar^2}{2A'a^2}. \quad (5.27)$$

It is a negative quantity. Such electrons behave abnormally in an external field set up in a crystal: they are accelerated in the direction opposite to the acting force. The absolute value is, as before, determined by the width of the energy band: m'_n is the smaller the wider the band is.

Let us now find what is responsible for such a "strange" behaviour of the electron in a crystal.

In case of a free electron, all the work W performed by an external force \mathbf{F} is spent to increase the kinetic energy of the electron's translational motion:

$$W = E_k = \frac{mv^2}{2} = \frac{\hbar^2 k^2}{2m}.$$

Differentiating E_k twice with respect to k , we obtain, $d^2E_k/dk^2 = \hbar^2/m$. Substituting into Eq. (5.25), we find that $m_{\text{eff}} = m$. Hence, the effective mass of a free electron is simply its rest mass.

The situation may be quite different in a crystal where the electron has not only kinetic energy but potential energy as well. When the electron moves under

the action of an external force F , some of the work performed by this force may be transformed into kinetic energy E'_k , the rest being transformed into potential energy U , and $W = E'_k + U$. In this case, the result will be a smaller increase in kinetic energy and, consequently, in electron velocity than in the case of a free electron. The electron, figuratively speaking, gains weight and moves under the action of the force F with a smaller acceleration than a free electron would.

Should the entire work of the external force be transformed into potential energy U , that is $W = U$, there would be no increase in the kinetic energy and in the velocity of the electron and it would behave as a particle of infinite effective mass.

Finally, if in the course of the electron's motion, not only the entire work of the external force F , but some of the kinetic energy E'_k that the electron had initially, too shall be transformed into the potential energy, so that $U = W + E'_k$, then as the electron moves along the crystal its velocity shall diminish, it shall be accelerated, behaving as a particle with a *negative effective mass*. Just such is the behaviour of electrons occupying states close to the top of the conduction band.

However, a situation is possible in a crystal when in the course of the electron's motion under the action of an external force F , not only the entire work of this force but some of the electron's potential energy, say U' , shall be transformed into its kinetic energy, so that $E''_k = W + U'$. The E''_k and the velocity v of such an electron shall rise quicker than those of a free electron. It loses weight as compared with the free electron, so that its effective mass $m_{\text{eff}} < m$.

The aforesaid is illustrated in Figure 5.12, which shows the nature of the variation of the total energy of the electron $E(k)$, of its translational velocity $v(k)$, and of its effective mass m_{eff} with the rise in k from 0 to $\pm\pi/a$.

Close to the bottom of the band ($k = 0$), as long as the electron's energy $E(k)$ rises approximately in proportion to k^2 , the velocity of the electron's translational motion $v \propto dE/dk$ increases in proportion to m_n , the acceleration remains positive, and the effective mass $m_{\text{eff}} \propto (d^2E/dk^2)^{-1}$ retains its constant positive value m_n . At the point of inflection C of the curve $E(k)$, the second derivative d^2E/dk^2 vanishes and the first derivative reaches its maximum value. Therefore, as this point is approached, $m_{\text{eff}} \rightarrow \infty$ and $v \rightarrow v_{\text{max}}$. After the point of inflection, dE/dk starts to decrease causing a decrease in v ; hence, acceleration becomes negative, which, for the direction of the external force F remaining unchanged, is equivalent to a change in the sign of the effective mass from the positive to the negative. If the curvature of the curve $E(k)$, which is proportional to d^2E/dk^2 , also changes, this would lead to a change in the absolute value of $m_{\text{eff}} \propto (d^2E/dk^2)^{-1}$. Near the top of the band $E(k)$ again becomes a quadratic function of k , and the effective mass assumes a constant negative value m'_n .

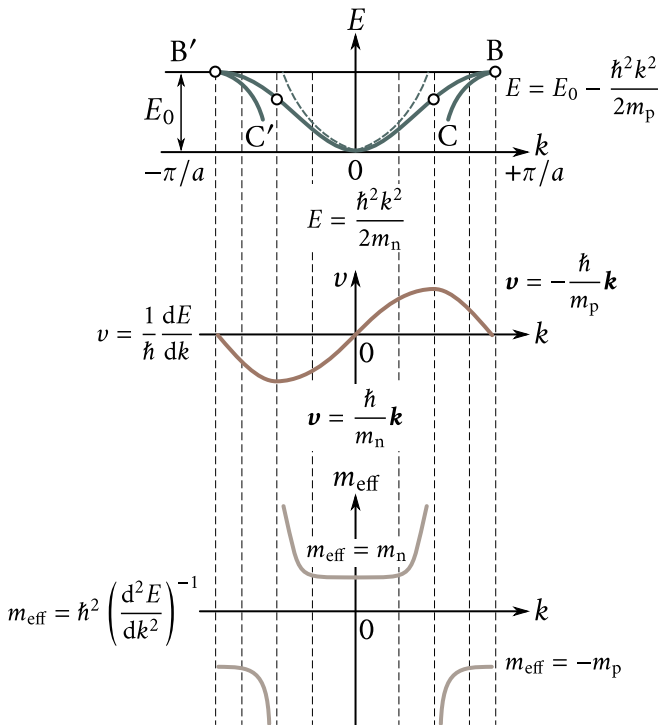


Figure 5.12: Dependence of electron energy E , group velocity v , and effective electron mass m_{eff} on k .

§ 42. Occupation of bands by electrons. Conductors, dielectrics, and semiconductors

Each energy band contains a limited number of energy levels. In compliance with the Pauli exclusion principle each level may be occupied by no more than two electrons. With the limited number of electrons in a solid only the lower energy bands will be filled. According to the nature of band occupation by electrons all solids can be classified into two large groups.

The *first* group includes bodies in which there is a partially filled band above the completely filled lower bands [Figure 5.13(a)]. Such bands are formed from partially filled atomic levels as, for instance, in the case of alkali metals. A partially filled band may also be the result of overlapping of filled and empty or partially filled bands, as is the case with beryllium and alkali-earth metals [Figure 5.13(b)]. A partially filled band is a feature of metals.

The *second* group includes bodies with empty bands above completely filled

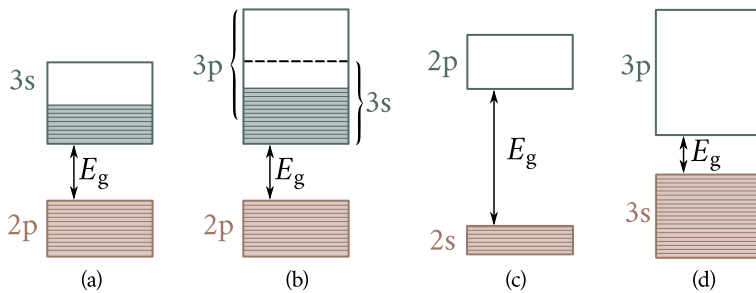


Figure 5.13: Occupation of bands by electrons: (a, b)—there is a partially filled band above the filled band; (c, d)—there is an empty band above the filled band.

ones [Figure 5.13(c,d)]. Typical examples of such bodies are the elements of Group IV of the Mendeleev periodic table—carbon in the diamond modification, silicon, germanium and gray tin, which has the structure of diamond. This group also includes many chemical compounds—metal oxides, nitrides, carbides, halides of alkali metals, etc.

According to the band theory of solids, the electrons of the outermost energy bands have practically the same freedom of movement, no matter whether the solid is a metal or a dielectric. The motion takes place by means of tunneling from atom to atom. Despite this fact, there is a difference of many orders of magnitude in the electric properties, in particular in the electrical conductivity, of the bodies of both types: in metals $\sigma \approx 10^7 (\Omega \text{ m})^{-1}$, and in good dielectrics $\sigma \approx 10^{-11} (\Omega \text{ m})^{-1}$.

To gain insight into the mechanism of electrical conductivity of solids let us discuss the behaviour of the electrons of partially and completely filled energy bands in an external field.

Set up an external field \mathcal{E} in the crystal. This field acts on every electron with a force $\mathbf{F} = -q\mathcal{E}$ that tends to distort the symmetry in the velocity distribution of the electrons, so that those moving against the force are decelerated and those moving in the direction of the force are accelerated. Since such acceleration and deceleration inevitably entail a change in the electron's energy, this is equivalent to the electron's transition to states with higher or lower energies. Such transitions may, evidently, take place only if there are unoccupied states inside the band to which the electrons belong, that is, if the band is not completely filled. In this case, even a weak electric field is capable of imparting to the electrons the necessary additional momentum that will take them to nearby free levels. A prevailing motion of the electrons in the direction opposite to that of the field will be set up in the solid resulting in an electric current. Such solids should be good conductors, which is actually the case.

Now let us imagine that the valence band of the crystal is completely filled and is separated from the nearby empty band by a wide energy gap E_g [Figure 5.13(c)]. An external field applied to such a crystal is incapable of changing the nature of electron motion in the valence band because it is unable to lift the electrons to the empty band lying above it. Inside the valence band, which has no free levels, the field may only cause the electrons to change places and this does not distort the symmetry of the electron distribution over velocities. Therefore, in such solids an external field is incapable of inducing directional motion of the electrons, that is, an electric current, and the electrical conductivity of such solids should be practically zero.

Hence, for a body to have high electrical conductivity it must have in its energy spectrum some energy bands only partially filled with electrons, as is the case with the typical metals [Figure 5.13(a,b)]. The absence of such bands in solids belonging to the second group makes them *nonconductors* despite the fact that they contain electrons weakly bonded to individual atoms.

The solids of the second group are conventionally subdivided into dielectrics and semiconductors according to the width of the forbidden band. *Dielectrics* include solids with a relatively wide forbidden band. For typical dielectrics, $E_g > 3$ eV. For diamond, $E_g = 5.2$ eV; for boron nitride, $E_g = 4.6$ eV; for Al_2O_3 , $E_g = 7$ eV; etc.

Semiconductors include solids with a relatively narrow forbidden band [Figure 5.13(d)]. For typical semiconductors, $E_g \lesssim 1$ eV. Thus, for germanium, $E_g = 0.66$ eV; for silicon, $E_g = 1.08$ eV; for indium antimonide, $E_g = 0.17$ eV; for gallium arsenide, $E_g = 1.43$ eV; etc.

Let us consider this class of solids in more detail.

§ 43. Intrinsic semiconductors. The concept of a hole

Intrinsic semiconductors. Semiconductors containing a negligible amount of electro-active defects (chemical and crystallographic) are termed *intrinsic semiconductors*. They include some pure chemical elements (germanium, silicon, selenium, tellurium, etc.) and numerous chemical compounds such as gallium arsenide (GaAs), indium arsenide (InAs), indium antimonide (InSb), silicon carbide (SiC), etc.

Figure 5.14(a) shows a simplified schematic diagram of an intrinsic semiconductor. At absolute zero, its valence band is completely filled and the conduction band, which is a distance E_g above the valence band, is empty. For this reason, at absolute zero, the intrinsic semiconductor, same as a dielectric, has zero conductivity.

However, as the temperature increases the electrons of the valence band be-

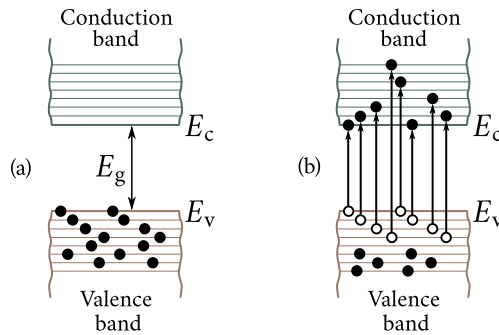


Figure 5.14: Intrinsic semiconductor: (a)—at absolute zero the valence band is completely filled by electrons and the conduction band is completely empty; (b)—at temperatures above absolute zero part of the electrons from the valence band are excited to the conduction band; holes appear in the valence band and free electrons in the conduction band (white circles denote holes and black circles denote electrons); E_c is the bottom of conduction band and E_v is the top of valence band.

come excited and some of them receive enough energy to surmount the forbidden band and go over to the conduction band [Figure 5.14(b)]. This results in free conduction electrons appearing in the conduction band and in free electron levels capable of accepting valence band electrons appearing in the valence band. When an external field is applied to such a crystal, a directional motion of the electrons of the conduction and the valence bands is established, resulting in the appearance of an electric current. The crystal becomes conducting.

The narrower the forbidden band and the higher the crystal's temperature the greater the number of electrons going over to the conduction band and, correspondingly, the greater the crystal's electrical conductivity. For instance, for germanium with $E_g = 0.66$ eV, the concentration of the electron gas in the conduction band already at room temperature is as high as $n_i \approx 10^{19} \text{ m}^{-3}$ and specific resistance is as low as $\rho_i \approx 0.48 (\Omega \text{ m})^{-1}$. At the same time, for diamond with $E_g = 5.2$ eV, n_i at room temperature is only about 10^4 m^{-3} and $\rho_i \approx 10^8 (\Omega \text{ m})^{-1}$. However, already at $T = 600 \text{ K}$, the electron gas concentration in diamond increases by many orders of magnitude and its specific resistance becomes as low as that of germanium at room temperature.

Two important conclusions may be drawn from the above.

- (1) The electrical conductivity of an intrinsic semiconductor is an excited conductivity: it appears only as a result of the action of some external factor capable of imparting sufficient energy to the electrons of the valence band to move them over to the conduction band. Such factors may be heating of the semiconductor and irradiation with light or with ionizing radiation.

- (2) The division of materials into semiconductors and dielectrics is essentially a matter of convention. Diamond—a dielectric at room temperature—exhibits a noticeable conductivity at higher temperatures and may also be considered to be a semiconductor. Materials with ever increasing forbidden band widths are now being used as semiconductors, gradually making the division into semiconductors and dielectrics irrelevant.

The concept of a hole. Let us now discuss in more detail the behaviour of the electrons in the valence band in which as a result of transition of some of the electrons to the conduction band some free levels have appeared [Figure 5.14(b)].

Now, the electrons of the valence band acted upon by an external field can go over to the free levels and establish an electric current in the crystal. Let us find the instantaneous value of this current.

The current established by one electron moving with a velocity \mathbf{v}_i is

$$\mathbf{I}_i = -q\mathbf{v}_i.$$

The total instantaneous current established by all the electrons of the valence band is

$$\mathbf{I}_t = -q \sum_i \mathbf{v}_i.$$

where the sum is over all the states occupied by electrons.

For a band completely filled with electrons, $\mathbf{I}_t = 0$, since there is an electron with the velocity \mathbf{v}_i to correspond to every electron with the velocity $-\mathbf{v}_i$.

Now let us imagine that all the states in the valence band except one with the velocity \mathbf{v}_s are filled. The total current in such a band will be

$$\mathbf{I} = -q \sum_{i \neq s} \mathbf{v}_i = -q \sum_i \mathbf{v}_i + q\mathbf{v}_s.$$

Since the first term in the right-hand side is zero,

$$\mathbf{I} = q\mathbf{v}_s. \quad (5.28)$$

Hence, the total current of all the electrons in a valence band with one vacant state is equivalent to a current set up by the motion of one particle with a positive charge q occupying the respective state. Such a fictitious particle is called a *hole*. If we attribute to the hole a positive charge $+q$ numerically equal to the electron charge, we should also attribute to it a positive effective mass m_p numerically equal to the negative effective mass of the electron m'_n , which initially occupied that state close to the top of the valence band, since only in this case will the current established by holes coincide both in magnitude and in direction with the current established by the electrons of the almost completely filled valence band.

Table 5.2 presents the room temperature electrophysical properties and char-

acteristics of the band pattern of three typical intrinsic semiconductors—silicon, germanium, and indium antimonide.

We see that a reduction in the forbidden band width is followed by a drastic rise in the concentration of free charge carriers in the semiconductor and a drop in its specific resistance. It may be seen from the two last columns of the table that the effective mass of the charge carriers may be much smaller than the electron rest mass.

§ 44. Impurity semiconductors

Semiconductors no matter how pure they are always contain some impurity atoms, which create their own energy levels termed *impurity levels*. Those levels may occupy positions both inside the allowed and the forbidden bands of the semiconductor at various distances from the top of the valence band and from the bottom of the conduction band. Frequently the impurities are introduced intentionally to impart specific properties to the semiconductor. Let us consider the main types of impurity levels.

Donor levels. Suppose that some germanium atoms in a germanium crystal are replaced by pentavalent arsenic atoms. Germanium has a diamond type lattice in which every atom is surrounded by four nearest neighbours bound to it by valence forces [Figure 5.15(a)]. To establish bonds with those neighbours the arsenic atom uses four valence electrons; the fifth electron takes no part in the bonding. It continues to move in the field of the arsenic ion, where the field is reduced in germanium by a factor of $\epsilon = 16$ (ϵ is the relative permittivity of germanium). Because of a weaker field, the radius of the electron's orbit increases 16-fold (as compared with that in an isolated atom) and its bond energy with the arsenic atom decreases about $\epsilon^2 \approx 256$ times, becoming equal to $E_d \approx 0.01$ eV. When this energy is imparted to the electron, the electron leaves the atom and is now free to move in the

Table 5.2

Semiconductor	E_g , (eV)	n_i , (m^{-3})	ρ_i , ($\Omega \text{ m}$)	Effective mass	
				m_n	m_p
Silicon	1.12	$\sim 10^{16}$	2×10^3	$1.08m$	$0.37m$
Germanium	0.66	3×10^{19}	0.48	$0.56m$	$0.59m$
Indium antimonide	0.17	1.6×10^{22}	6×10^{-5}	$0.015m$	$0.18m$

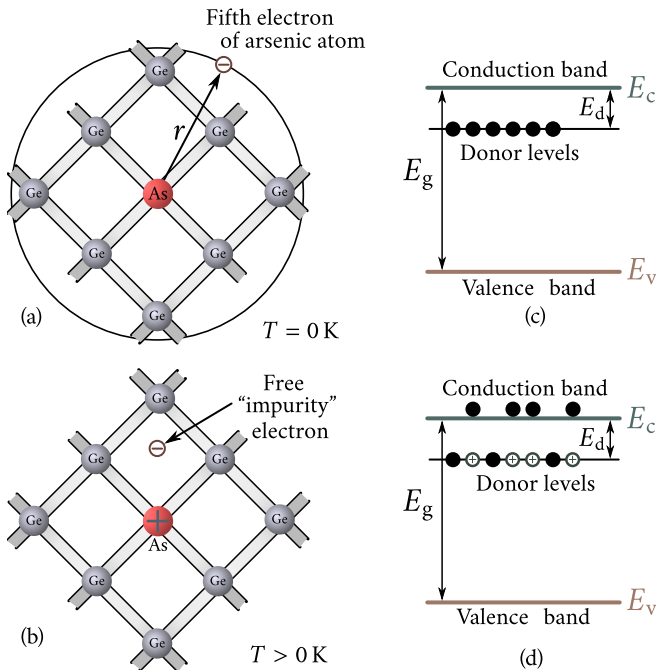


Figure 5.15: Charge carrier excitation in an n-type semiconductor: (a)—at $T = 0 \text{ K}$, the atoms of pentavalent arsenic in the germanium lattice are in a nonionized state; (b)—ionization of arsenic atoms and generation of conduction electrons at $T > 0 \text{ K}$; (c)—energy levels of one of the five electrons of every arsenic atom are donor levels; (d)—electron transition from a donor level to the conduction band at $T > 0 \text{ K}$.

germanium lattice thereby becoming a conduction electron [Figure 5.15(b)].

In terms of band theory this process may be described as follows. The energy levels of the fifth electron of the arsenic atom occupy positions between the valence band and the conduction band [Figure 5.15(c)]. Those positions are directly under the bottom of the conduction band at a distance of $E_d \approx 0.01 \text{ eV}$ from it. When an electron occupying such an impurity level receives additional energy greater than E_d , it goes over to the conduction band [Figure 5.15(d)]. The remaining positive charge (a “hole”) is localized on the immobile arsenic atom and does not take part in electrical conductivity.

The impurities which supply electrons are termed *donors* and the energy levels of those impurities *donor levels*. The semiconductors doped with donor impurities are termed *n-type semiconductors*.

Acceptor levels. Let us suppose now that some of the germanium atoms in the germanium lattice are replaced by trivalent indium atoms [Figure 5.16(a)]. The

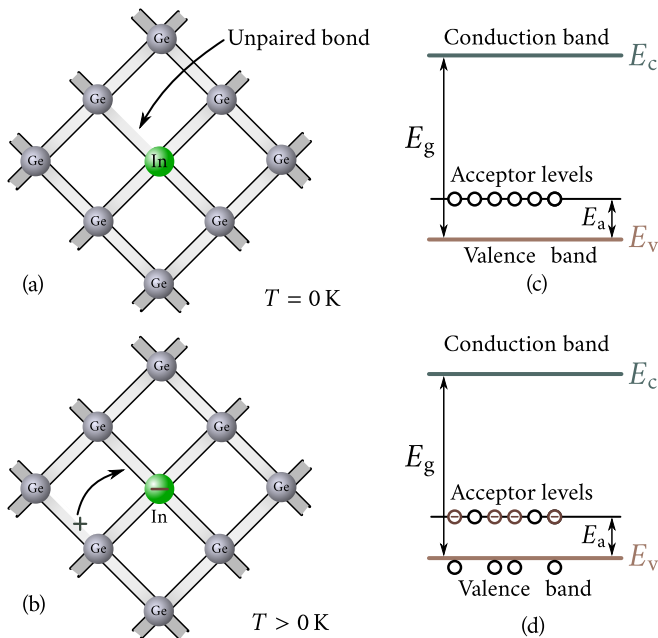


Figure 5.16: Charge carrier excitation in a p-type semiconductor: (a)—atoms of trivalent indium in the germanium lattice at $T = 0\text{ K}$ (the fourth bond of the indium atom is unpaired); (b)—at $T > 0\text{ K}$, the electrons can go over to unpaired bonds of impurity atoms creating an indium ion and a vacant level (hole) in the valence band of germanium; (c)—energy levels of unpaired bonds of indium atoms are acceptor levels; (d)—electron transition from the valence band to an acceptor level at $T > 0\text{ K}$ result in the generation of holes in this band.

indium atom lacks one electron to establish bonds with all the four nearest neighbours. It can “borrow” this electron from a germanium atom. Calculations show that the necessary energy is of the order of $E_a \approx 0.01\text{ eV}$. The ruptured bond corresponds to a hole [Figure 5.16(b)] since it results in the formation of a vacant state in the valence band of germanium.

Figure 5.16(c) shows the band pattern of germanium doped with indium. Directly above the valence band at a distance of $E_a \approx 0.01\text{ eV}$ away from it there are some empty levels of the indium atoms. Those levels are so close to the valence band that already at relatively low temperatures, some electrons from the valence band go over to the impurity levels [Figure 5.16(d)]. They establish bonds with the indium atoms and lose their ability to move in the germanium lattice playing no part in the conductivity. Only the holes created in the valence band act as charge carriers.

The impurities that trap electrons from the valence band are termed *acceptors*

and the energy levels of such impurities *acceptor levels*. The semiconductors doped with such impurities are termed *p-type semiconductors*.

§ 45. Position of the Fermi level and free carrier concentration in semiconductors

Dependence of free carrier concentration on the position of the Fermi level.

One of the main parameters of the gas of free carriers in a semiconductor is its chemical potential, μ . As applied to the electron and hole gases the usual term for it is the *Fermi level*.

As we have ascertained in Chapter 3, the Fermi level in metals is the last occupied level in the conduction band (see Figure 3.4): at absolute zero, all levels below the Fermi level are occupied by electrons and all levels above the Fermi level are empty. The concentration of the electron gas in metals is comparable, as regards its order of magnitude, to the number of states in the conduction band; because of this the gas is degenerate and the distribution of the electrons over the states is described by the Fermi-Dirac quantum statistics. The electron concentration of such a gas is practically independent of temperature.

In the intrinsic and low-doped impurity semiconductors the electron (the hole) gas is nondegenerate and the distribution of electrons over the states is described by the Maxwell-Boltzmann classical statistics. For such semiconductors the free carrier concentration is dependent on the position of the Fermi level and on temperature. Let us find this dependence.

Figure 5.17 shows the band pattern of a nondegenerate semiconductor. At some temperature T other than absolute zero, there are some electrons in the conduction band of such a semiconductor and some holes in its valence band. Denote their concentrations by n and p , respectively. Take as the zero energy level the bottom of the conduction band. Choose a small energy interval dE lying between E and $E + dE$ close to the bottom of the conduction band. Since the electron gas in a semiconductor is a nondegenerate gas, the number of electrons dn in the energy interval dE (per unit volume of semiconductor) may be calculated with the aid of Eq. (3.28):

$$dn = \frac{4\pi}{h^3} (2m_n)^{3/2} e^{\mu/(k_B T)} e^{-E/(k_B T)} E^{1/2} dE. \quad (5.29)$$

In nondegenerate semiconductors μ is negative [see Eq. (3.48)]. This means that the Fermi level in such semiconductors is below the bottom of the conduction band, as shown in Figure 5.17. Denote the distance from the bottom of the conduction band to the Fermi level and from the Fermi level to the top of the valence band

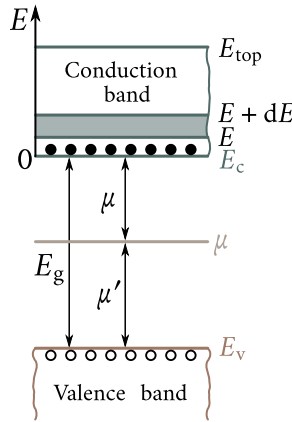


Figure 5.17: Band pattern of a nondegenerate semiconductor: E_c is the bottom of conduction band, E_v the top of valence band, μ the Fermi level, and E_g the forbidden band width.

by μ and μ' , respectively. Evidently

$$\mu + \mu' = -E_g, \quad \text{or} \quad \mu' = -(E_g + \mu) \quad (5.30)$$

where E_g is the width of the forbidden band of the semiconductor.

To obtain the total number of electrons in the conduction band at temperature T we integrate Eq. (5.29) over the energy values corresponding to the conduction band, that is, from 0 to E_{top} :

$$n = 4\pi \left(\frac{2m_n}{h^2} \right)^{3/2} e^{\mu/(k_B T)} \int_0^{E_{\text{top}}} e^{-E/(k_B T)} E^{1/2} dE.$$

The function $e^{-E/(k_B T)}$ decreases very rapidly as E grows; therefore, it is permissible to substitute infinity for the upper limit to obtain

$$n = 4\pi \left(\frac{2m_n}{h^2} \right)^{3/2} e^{\mu/(k_B T)} \int_0^{\infty} e^{-E/(k_B T)} E^{1/2} dE.$$

Evaluation of this integral yields

$$n = 2 \left(\frac{2\pi m_n k_B T}{h^2} \right)^{3/2} e^{\mu/(k_B T)}. \quad (5.31)$$

A similar calculation carried out for holes generated in the valence band yields the expression

$$p = 2 \left(\frac{2\pi m_p k_B T}{h^2} \right)^{3/2} e^{-(E_g + \mu)/(k_B T)} = 2 \left(\frac{2\pi m_p k_B T}{h^2} \right)^{3/2} e^{\mu'/(k_B T)} \quad (5.32)$$

where m_p is the effective mass of the hole.

It follows from Eqs. (5.31) and (5.32) that the concentration of free charge carriers

in a band is determined by the distance from the boundary of this band to the Fermi level: the greater this distance the smaller the carrier concentration (since $\mu < 0$ and $\mu' < 0$).

According to Eqs. (5.31) and (5.32) the product of n and p for any nondegenerate semiconductor is

$$np = 4 \left(\frac{2\pi k_B T}{h^2} \right)^3 (m_n m_p)^{3/2} e^{-E_g/(k_B T)}. \quad (5.33)$$

Formula (5.33) shows that for a definite temperature the product of the electron and hole concentrations is a constant for the respective semiconductor. This is an expression of the law of mass action as applied to the free carrier gas in semiconductors.

Let us now discuss separately the position of the Fermi level and the free carrier concentration in intrinsic and impurity semiconductors.

Position of Fermi level and carrier concentration in intrinsic semiconductors. In intrinsic semiconductors the concentration of electrons in the conduction band, n_i , is equal to that of holes in the valence band, p_i :

$$n_i = p_i \quad (5.34)$$

since every electron that goes over to the conduction band leaves behind a hole in the valence band. Equating right-hand sides of Eqs. (5.31) and (5.32), we obtain

$$2 \left(\frac{2\pi m_n k_B T}{h^2} \right)^{3/2} e^{\mu/(k_B T)} = 2 \left(\frac{2\pi m_p k_B T}{h^2} \right)^{3/2} e^{-(E_g + \mu)/(k_B T)}.$$

Solving this equation for μ , we obtain

$$\mu = -\frac{E_g}{2} + \frac{3}{4} k_B T \ln \left(\frac{m_p}{m_n} \right). \quad (5.35)$$

This relation determines the position of the Fermi level in intrinsic semiconductors. At absolute zero ($T = 0$ K)

$$\mu = -\frac{E_g}{2} \quad (5.36)$$

that is, the position of the Fermi level is exactly the middle of the forbidden band (Figure 5.18). As the temperature rises the Fermi level shifts upwards towards the bottom of the conduction band if $m_p > m_n$, and downwards towards the top of the valence band if $m_p < m_n$. In many cases, however, the shift is so small that the Fermi level in intrinsic semiconductors can be considered to be always in the middle of the forbidden band. Substituting μ from Eq. (5.35) into (5.31) and (5.32), we obtain

$$n_i = p_i = 2 \left(\frac{2\pi \sqrt{m_n m_p} k_B T}{h^2} \right)^{3/2} e^{-E_g/(2k_B T)}. \quad (5.37)$$

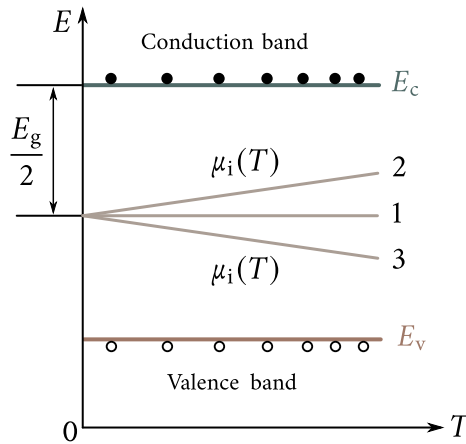


Figure 5.18: Position of Fermi level in an intrinsic semiconductor at various temperatures: at $T = 0$ K the Fermi level is in the middle of the forbidden band; as temperature rises the Fermi level does not change its position if $m_n = m_p$ (straight line 1), shifts upwards if $m_n < m_p$ (straight line 2), and shifts downwards if $m_n > m_p$ (straight line 3).

It follows from Eq. (5.37) that the equilibrium carrier concentration in an intrinsic semiconductor is determined by the width of the forbidden band and the temperature of the semiconductor, and the dependence on T and E_g is very strong. For instance, at room temperature a decrease in E_g from 1.12 eV (silicon) to 0.08 eV (gray tin) results in an increase in nine orders of magnitude of n . An increase in the temperature of germanium from 100 K to 600 K increases n by 17 orders of magnitude.

Using Eq. (5.37), we may rewrite the law of mass action (5.33) as

$$np = n_i^2. \quad (5.38)$$

Position of Fermi level and carrier concentration in impurity semiconductors. Figure 5.19 shows the change in the Fermi level position with the increase in temperature for (a) n- and (b) p-type semiconductors.

The low temperature range. At low temperatures the average energy of lattice thermal vibrations, $k_B T$, is much less than the width of the forbidden band, E_g , and because of that, the vibrations are incapable of providing sufficient excitation of the electrons of the valence band to shift them to the conduction band. But this energy is enough to excite and shift to the conduction band the electrons occupying the donor levels E_d (Figure 5.20) and to the valence band the holes occupying the acceptor levels E_a (Figure 5.21), since this requires an energy 100 times less than E_g . Therefore, at low temperatures practically only the “impurity” charge carriers—electrons in the n-type semiconductors and holes in the p-type—are excited

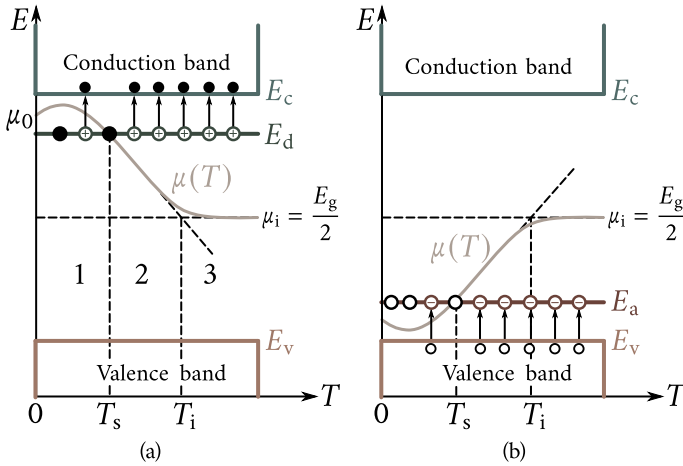


Figure 5.19: Variation of Fermi level position with temperature: (a)—in n-type semiconductors; (b)—in p-type semiconductors (T_s is the saturation temperature of impurity levels and T_i the temperature of transition to intrinsic conductivity).

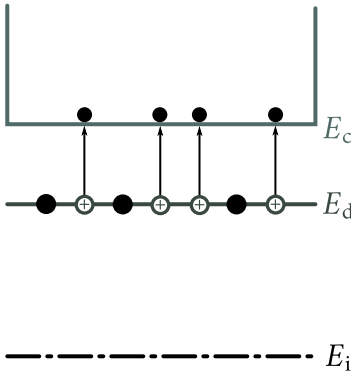


Figure 5.20: Excitation of the electrons occupying a donor level and their transition to the conduction band.

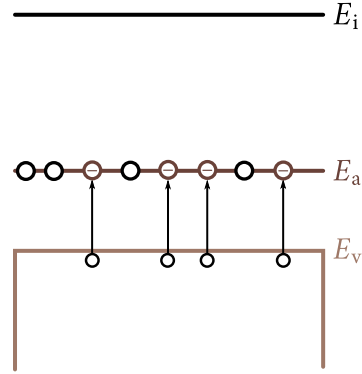


Figure 5.21: Excitation of the electrons occupying the valence band and their transition to an acceptor level.

in impurity semiconductors.

Calculations show that the position of the Fermi level inside this temperature range is

$$\mu = -\frac{E_d}{2} + \frac{k_B T}{2} \ln \left[\frac{N_d h^3}{2 (2\pi m_n k_B T)^{3/2}} \right] \quad (5.39)$$

for the n-type semiconductors, and

$$\mu' = -\frac{E_a}{2} + \frac{k_B T}{2} \ln \left[\frac{N_a h^3}{2 (2\pi m_p k_B T)^{3/2}} \right] \quad (5.40)$$

for the p-type semiconductors, N_d and N_a being the concentrations of the donors and acceptors. Graphs of the temperature dependence of μ corresponding to those functions are presented in Figure 5.19(a, b).

Substituting μ and μ' from (5.39) and (5.40) into (5.31) and (5.32), respectively, we obtain the following expressions for the concentrations:

$$n = \sqrt{2N_d} \left(\frac{2\pi m_n k_B T}{h^2} \right)^{3/2} e^{-E_d/(2k_B T)} \quad (5.41)$$

of electrons in the n-type semiconductors and

$$p = \sqrt{2N_a} \left(\frac{2\pi m_p k_B T}{h^2} \right)^{3/2} e^{-E_a/(2k_B T)} \quad (5.42)$$

of holes in the p-type semiconductors.

The impurity exhaustion range (extrinsic range). As temperature rises, the electron concentration in the conduction band increases and that on the donor levels decreases—the donor levels become exhausted. The behaviour of acceptor levels in p-type semiconductors is similar.

In case of complete exhaustion the electron concentration in the conduction band of an n-type semiconductor becomes practically equal to the concentration of donor impurity, N_d :

$$n \approx N_d \quad (5.43)$$

and the hole concentration in a p-type semiconductor, to that of acceptor impurity, N_a :

$$p \approx N_a. \quad (5.43')$$

The *exhaustion*, or *saturation*, temperature of the impurity levels, T_s , is the higher the higher the impurity's activation energy, E_d or E_a , and its concentration. For instance, for germanium with $N_d = 10^{22} \text{ m}^{-3}$ and $E_d = 0.01 \text{ eV}$ the saturation temperature is approximately 30 K.

The high temperature range. As the temperature is raised still higher the excitation of intrinsic carriers becomes more intense, the semiconductor increasingly approaching the state of an intrinsic semiconductor with the Fermi level approaching the position of that in an intrinsic semiconductor. Until the concentration of intrinsic carriers remains much less than N_d ($n_i \ll N_d$), the total concentration $n = n_i + N_d$ remains practically constant and equal approximately to N_d .

However, at sufficiently high temperatures the intrinsic carrier concentration

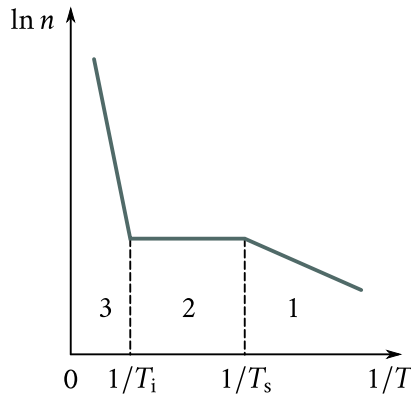


Figure 5.22: Temperature dependence of electron concentration in n-type semiconductors: 1—impurity conductivity range, 2—impurity exhaustion range, 3—intrinsic conductivity range.

may not only become equal to N_d but may substantially exceed it ($n_i \gg N_d$). In this case $n = n_i + N_d$ is approximately n_i and marks the transition to intrinsic conductivity. The temperature T_i of such transition is the higher the greater is the width of the forbidden band and the impurity concentration. For germanium with $N_d = 10^{22} \text{ m}^{-3}$ this temperature is 450 K.

Above T_i , the Fermi level in an impurity semiconductor coincides with the Fermi level in an intrinsic semiconductor and is expressed by Eq. (5.35), its carrier concentration being identical to that of an intrinsic semiconductor at that temperature, as described by Eq. (5.37). Figure 5.22 shows schematically the dependence of the natural logarithm of the electron concentration in the conduction band of an n-type semiconductor on the reciprocal temperature. Three sections may be marked on the curve: 1, corresponding to impurity conduction; 2, corresponding to the impurity exhaustion range; and 3, corresponding to intrinsic conductivity.

Finally, it should be pointed out that in contrast to intrinsic semiconductors, in which both electrons and holes simultaneously take part in electrical conductivity, in impurity semiconductors the conductivity is mainly due to charge carriers of one sign: to electrons in the n-type semiconductors and to holes in the p-type. Such carriers are termed *majority carriers*.

Apart from them, a semiconductor always contains minority carriers as well: n-type semiconductors contain holes and p-type semiconductors, electrons. Equilibrium carrier concentrations may be conveniently denoted as follows: n_{n0} and p_{n0} are the concentrations of electrons (majority carriers) and holes (minority carriers) in n-type semiconductors, p_{p0} and n_{p0} are the concentrations of holes (majority carriers) and electrons (minority carriers) in p-type semiconductors.

Using this notation, we may write the law of mass action (5.38) in the following form:

$$n_{n0}p_{n0} = n_i^2, \quad p_{p0}n_{p0} = n^2. \quad (5.44)$$

It follows from Eq. (5.44) that doping a semiconductor by an electrically active impurity, while increasing the majority carrier concentration, should inevitably decrease the minority carrier concentration so as to keep the product of those concentrations constant.

§ 46. Nonequilibrium carriers

As we already know, at all temperatures other than absolute zero a process of *free carrier excitation*, or *generation*, takes place in the semiconductor. Should this be the only process taking place, the carrier concentration would continuously grow with time. However, there is a simultaneous process of *free carrier recombination*. The essence of this process is that when a free electron meets a hole it may occupy it, the result being annihilation of a pair of carriers.

At any temperature, an equilibrium is established between the processes of thermal carrier generation and recombination characterized by appropriate equilibrium carrier concentrations. Such carriers are termed *equilibrium carriers*. The law of mass action discussed in the previous section is applicable only to them.

Besides thermal excitation, there are other methods of free carrier generation in semiconductors: by light, by ionizing particles, by injection through a contact, and others. Such factors result in the appearance of additional free carriers, *excess carriers*, as compared with the equilibrium carrier concentration. Another term for them is *nonequilibrium carriers*. Denote the concentrations of such carriers by Δn and Δp , respectively. Then the total carrier concentration will be

$$n = n_0 + \Delta n, \quad p = p_0 + \Delta p \quad (5.45)$$

where n_0 and p_0 are the equilibrium carrier concentrations.

Every nonequilibrium carrier having been born in the semiconductor “lives” a limited time before recombining, the time being different for different carriers. For this reason, an average carrier lifetime τ is introduced, with the notation τ_n for electrons and τ_p for holes.

The carrier generation process is characterized by the *generation rate* g , which expresses the number of carriers (or carrier pairs) generated in a unit volume of the semiconductor per second.

The recombination process is characterized by the *recombination rate* R , which is equal to the number of carriers (carrier pairs) recombining in a unit volume of

the semiconductor per second. For electrons

$$R_n = -\frac{dn}{dt} = -\frac{d(\Delta n)}{dt} \quad (5.46)$$

and for holes

$$R_p = -\frac{dp}{dt} = -\frac{d(\Delta p)}{dt} \quad (5.47)$$

where n and p are the total concentrations of electrons and holes, respectively, at a given moment of time; Δn and Δp are the respective excess concentrations at the same moment; and the minus signs point to the fact that recombination results in a decrease in carrier concentrations.

Suppose that light generates excess carriers in a semiconductor whose concentrations are $\Delta n_0 = \Delta p_0$. After the light is turned off, those carriers will recombine and their concentrations shall gradually diminish. Since every excess carrier, for instance, an electron, lives on the average τ_n , their recombination rate will be $\Delta n/\tau$ per second, where Δn is the excess carrier concentration at the moment. Therefore, the recombination rate is

$$R_n = -\frac{d(\Delta n)}{dt} = \frac{\Delta n}{\tau_n}. \quad (5.48)$$

A similar relation holds for holes:

$$R_p = -\frac{d(\Delta p)}{dt} = \frac{\Delta p}{\tau_p}. \quad (5.49)$$

Integrating the two equations, we obtain

$$\Delta n = \Delta n_0 e^{-t/\tau_n}, \quad \Delta p = \Delta p_0 e^{-t/\tau_p}. \quad (5.50)$$

It follows from Eq. (5.50) that $\Delta n = \Delta n_0/e$ and $\Delta p = \Delta p_0/e$ for $t = \tau$. Hence, the average excess carrier lifetime is a time interval during which the carrier concentration due to recombination decreases $e = 2.73$ times.

Free charge carriers diffuse in the volume of the semiconductor and during their lifetime τ cover an average distance L termed *carrier diffusion length*. Calculations show that L depends on τ in the following manner:

$$L = \sqrt{Dt} \quad (5.51)$$

where D is the *carrier diffusion coefficient*, related to their mobility u by the Einstein relation

$$D = \frac{k_B T u}{q} \quad (5.52)$$

where q is the electron charge.

The transition of the electron from the conduction to the valence band may take place directly across the whole forbidden band E_g , as shown by arrow 1 in Figure 5.23, or indirectly, first to the impurity level E_{im} (arrow 2) and then from

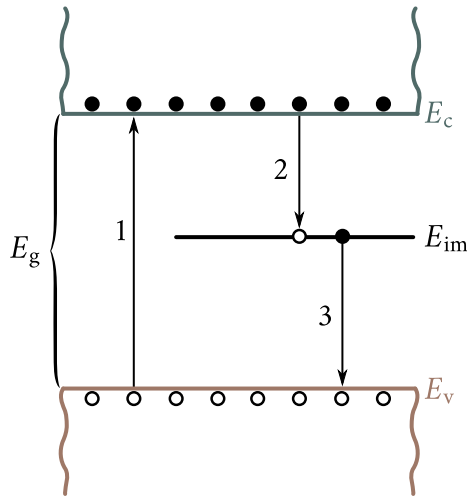


Figure 5.23: Excess charge carrier recombination in semiconductors: 1—direct recombination, 2 and 3—recombination via impurity level.

this level to the valence band (arrow 3). Recombination of the first type is termed *direct recombination* and of the second type *recombination via an impurity level*.

In both types of recombination the same energy E_g is liberated. The only difference is that in the first case this energy is liberated in one act and in the second in two acts corresponding to the transitions 2 and 3.

The energy may be liberated in the form of a light quantum $h\nu$ or in the form of heat (phonons). In the first instance the recombination is termed *radiative* and in the second *nonradiative*. Calculation and experiment show that direct recombination plays an essential part in semiconductors with a narrow forbidden band at relatively high temperatures (room temperature and above). The principal recombination mechanism in wide forbidden band semiconductors is nonradiative recombination via impurity, or defect, levels. However, under appropriate conditions a relatively high level of radiative recombination may be attained even in such semiconductors. A favourable factor is, for instance, the increase in excess carrier concentration and in some cases higher doping. A remarkable material in this respect is gallium arsenide (GaAs) in which, given favourable conditions, radiative recombination may constitute as high as 50% or higher of the total. For this reason, gallium arsenide is at present the principal material for making luminescent diodes and semiconductor lasers, which find wide practical use.

Chapter 6

Electrical Conductivity of Solids

§ 47. Equilibrium state of electron gas in a conductor in the absence of an electric field

In the absence of an electric field, the electron gas in a conductor is in an equilibrium state described by equilibrium distribution functions. For a degenerate gas the appropriate function is the Fermi-Dirac distribution function [Figure 6.1(a)] and for a nondegenerate gas the Maxwell-Boltzmann distribution function [Figure 6.1(b)].

It may be seen from Figure 6.1 that the graphs of those functions are symmetric about the axis of ordinates. This points to the fact that the number of electrons in a conductor moving in the opposite directions is always the same and their average velocity in any direction is zero. This explains the fact that there is no electric current in a conductor (in the absence of a field), no matter how many free electrons it contains.

The equilibrium of the electron gas is established as a result of the interaction of the electrons with the lattice defects, this interaction being accompanied by energy and momentum exchanges. Such defects include thermal vibrations of the lattice (phonons), lattice imperfections, and impurity atoms. The interaction results in electron scattering and their random motion in the conductor.

§ 48. Electron drift in an electric field

When an electric field \mathcal{E} is applied to a conductor, an electric current is established in it whose density according to Ohm's law is

$$\mathbf{i} = \sigma \mathcal{E}. \quad (6.1)$$

The proportionality factor σ is termed the *specific conductance* of the conductor. Its

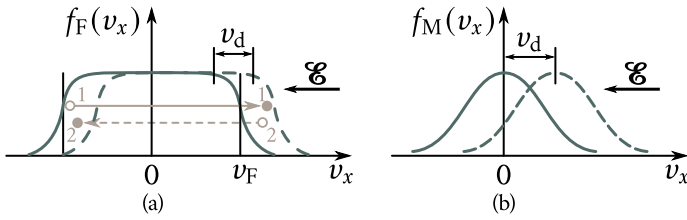


Figure 6.1: The Fermi-Dirac (a) and Maxwell-Boltzmann (b) distribution functions.

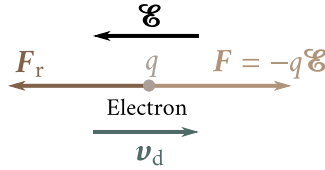


Figure 6.2: Forces acting on a free electron in a conductor in which an electric field \mathcal{E} has been established.

dimensions are $\Omega^{-1} \text{ cm}^{-1}$ or $\Omega^{-1} \text{ m}^{-1}$. Good conductors have $\sigma \approx 10^7 \Omega^{-1} \text{ m}^{-1}$ to $10^8 \Omega^{-1} \text{ m}^{-1}$; good dielectrics $10^{-12} \Omega^{-1} \text{ m}^{-1}$ to $10^{-14} \Omega^{-1} \text{ m}^{-1}$. Often, it is more convenient to use the *specific resistance*

$$\rho = \frac{1}{\sigma} \quad (6.2)$$

instead of specific conductance.

The specific resistance is measured in $\Omega \text{ m}$. For metals $\rho \approx 10^{-7} \Omega \text{ m}$ to $10^{-8} \Omega \text{ m}$; for dielectrics $\rho \approx 10^{12} \Omega \text{ m}$ to $10^{14} \Omega \text{ m}$.

A current flowing in the conductor is an indication of the fact that electrons acted upon by the field begin to move in a specific direction and that their distribution function experiences a change. Such directional motion is termed *drift* and the average velocity of this motion—*drift velocity* \mathbf{v}_d . Let us calculate it.

The force with which the field \mathcal{E} acts on an electron is $\mathbf{F} = -q\mathcal{E}$ (Figure 6.2). Acted upon by this field the electron should be accelerated and its velocity should grow continuously. But in its motion the electron collides with the lattice defects and as a result of scattering loses the velocity it gained in the field. The effect of the lattice may be formally reduced to the action of a resistance force \mathbf{F}_r , which hinders the electron in its motion through the lattice. This force is proportional to \mathbf{v}_d and is directed against it:

$$\mathbf{F}_r = -\frac{1}{\tau} m_n \mathbf{v}_d \quad (6.3)$$

with $1/\tau$ a proportionality factor whose physical meaning will be made clear subsequently, and m_n the electron's effective mass.

Using Eq. (6.3), we may write the equation of the directional motion of the electron in the lattice in the following form:

$$m_n \frac{d\mathbf{v}_d(t)}{dt} = -q\mathcal{E} - \frac{1}{\tau} m_n \mathbf{v}_d(t). \quad (6.3')$$

We see from Eq. (6.3') that, after the field had been applied, the velocity of the directional motion of the electrons shall rise and the electrons shall be accelerated until the resistance force F_r , which is proportional to $\mathbf{v}_d(t)$, shall become equal to the force \mathbf{F} with which the field acts on the electron. When those forces become equal, the resultant force acting on the electron and, accordingly, its acceleration, shall vanish.

From this moment the directional motion of the electron shall proceed at a constant velocity

$$\mathbf{v}_d = -\frac{q\mathcal{E}\tau}{m_n}. \quad (6.4)$$

Since the electron charge is negative, its drift is in the direction opposite to \mathcal{E} .

The ratio of the drift velocity to the field intensity is termed carrier mobility:

$$u = \frac{v_d}{\mathcal{E}} = \frac{q\tau}{m_n}. \quad (6.5)$$

For electrons $u_n < 0$, and for holes $u_p > 0$.

According to Eq. (6.4), the drift velocity in a field of constant intensity remains constant. This is possible only if the force $\mathbf{F} = -q\mathcal{E}$ with which the field acts on the electron is compensated by the force F_r . Would the opposite be the case the drift velocity would grow continuously and could become infinitely high even in weak fields. In this case, the electrical conductance would be infinite and the electrical resistance would vanish.

This would be the case if free electrons moved in an ideal regular lattice with a strictly periodic potential. The electron wave that describes the behaviour of the electron in such a lattice would propagate in it practically without attenuation, similar to a light wave propagating in an optically transparent medium.

The causes of a finite electrical resistance are various lattice imperfections, which result in the deformation of the lattice periodic potential and which serve as scattering centres for the electron waves, and attenuate the directional flux of the electrons in the same way as light waves are scattered and a ray of light is attenuated in an opaque medium.

§ 49. Relaxation time and mean free path

Let us now find the physical meaning of the factor τ .

Suppose that as soon as the velocity of the directional motion of the electrons

attains its stationary value v_d , the field \mathcal{E} is turned off. Because of the collisions of the electrons with the lattice defects this velocity will start to diminish and the electron gas will return to the state of equilibrium. Such processes leading to the establishment of equilibrium in a system that was previously put out of it are termed *relaxation processes*.

Setting $q\mathcal{E} = 0$ in Eq. (6.3') yields an equation describing the return of the electron gas to the equilibrium state, that is, the relaxation process

$$\frac{dv_d(t)}{dt} = -\frac{1}{\tau}v_d(t). \quad (6.6)$$

Integrating (6.6), we obtain

$$v_d(t) = v_d e^{-t/\tau} \quad (6.7)$$

where $v_d(t)$ is the velocity of the directional motion of the electrons, (t is the time after the field had been turned off).

It follows from Eq. (6.7) that, τ , *characterizes the rate at which the equilibrium state of the system is established*: the less τ is, the sooner the excited system will return to the state of equilibrium. The velocity of directional motion of the electrons during the time $t = \tau$ decreases $e \approx 2.7$ times. The time τ is termed *relaxation time*. For pure metals $\tau \approx 10^{-14}$ s.

The motion of the electrons in a crystal may conveniently be described with the aid of the concept of mean free path. By analogy with the kinetic theory of gases, one may presume that an electron in a crystal moves along a straight line until it meets a lattice defect and is scattered. The average distance, λ , that the electron travels between two consecutive scattering acts is taken as the *mean free path* of the electron.

Should the electron lose its directional velocity completely already after a single scattering act returning to the former state of random motion, its mean free path would simply be the product of its average velocity and the relaxation time τ , which in this case would simply be the free transit time of the electron:

$$\lambda = v\tau. \quad (6.8)$$

However, often it is not one but on the average ν collisions with scattering centres that are required to nullify the directional velocity completely. Only after ν collisions, do all traces of correlation between the initial and the final velocities of the electron disappear. The time during which the directional motion of the electron becomes randomized will, in this case, too be termed relaxation time. However, the mean path the electron travels during this time is no longer λ , but

$$l = \nu\lambda = v\tau. \quad (6.9)$$

The quantity l is termed *transport mean free path*.

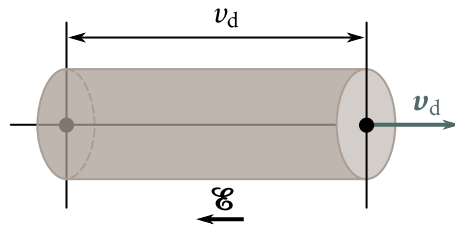


Figure 6.3: Calculating current density in a conductor.

It follows from Eq. (6.9) that

$$\tau = \frac{v\lambda}{v}. \quad (6.10)$$

The appearance of the drift of free charge carriers resulting in an electric current is an indication of the fact that, the field \mathcal{E} changes the distribution of free electrons over the states, that is, the form of the distribution function $f(E)$, since the equilibrium distribution function $f_0(E)$ cannot be the cause of the current. The dotted lines in Figures 6.1(a,b) show the graphs of the electron distribution functions after a constant drift velocity had been established. It may be seen from Figure 6.1 that the effect of the field \mathcal{E} on the electron distribution function over the states, is to shift the whole distribution by an amount $v_d = q\mathcal{E}\tau/m_n$ in the direction opposite to \mathcal{E} . Because of that shift, the distribution functions are no longer symmetric about the axis of ordinates and the average velocity of the electrons in the direction of the x axis is no longer zero (in the absence of the field this velocity was zero). It may be easily demonstrated that the average velocity will, in this case, be equal to the drift velocity v_d .

§ 50. Specific conductance of a conductor

Knowing the drift velocity of the electrons v_d , we can easily calculate the current density and the specific conductance of a conductor. To this end, imagine a cylinder with a unit base built inside a conductor with a side equal to v_d and directed along the direction of drift (Figure 6.3). All the electrons inside this cylinder will in one second pass through the base establishing a current with a density

$$\mathbf{i} = -qn\mathbf{v}_d = qnu\mathcal{E}. \quad (6.11)$$

Comparing Eq. (6.11) with (6.1), we obtain

$$\sigma = qnu. \quad (6.12)$$

Substituting u from Eq. (6.5) and τ from Eq. (6.10), we obtain

$$\sigma = \frac{nq^2}{m_n} \tau = \frac{nq^2}{m_n} \frac{\nu \bar{\lambda}}{\bar{v}}. \quad (6.13)$$

§ 51. Electrical conductivity of nondegenerate and degenerate gases

Up to now, we did not distinguish between the nondegenerate and degenerate electron gases. Let us now try and find how the state of an electron gas affects its electrical conductivity. To this end, we shall discuss in more detail the conductivity mechanism of the nondegenerate and the degenerate gases in more detail.

Nongenerate gas. In the case of a nondegenerate gas, the occupancy of the conduction band by electrons is so small that they practically never come so close together that their behaviour is limited by the Pauli exclusion principle. The electrons are perfectly free in the sense that the motion of any one of them is not noticeably affected by the others. Therefore, all the conduction electrons of a nondegenerate gas play an independent part in the electric current and in the electrical conductivity of the conductor. For this reason, formulae (6.5) and (6.13) for the electrical conductivity of the nondegenerate gas and for the electron mobility should include the mean free path $\bar{\lambda}$, the average number of collisions $\bar{\nu}$, the average velocity of motion \bar{v} , and the average relaxation time $\bar{\tau}$ of all the free electrons obtained by averaging over the ensemble as a whole.

Taking this into account, we can write the expressions for the electron mobility and for the specific conductance of a nondegenerate gas in the following form:

$$u = \frac{q\bar{\tau}}{m_n} = \frac{q}{m_n} \frac{\bar{\lambda}\bar{\nu}}{\bar{v}}, \quad (6.5')$$

$$\sigma = \frac{nq^2}{m_n} \bar{\tau} = \frac{nq^2}{m_n} \frac{\bar{\lambda}\bar{\nu}}{\bar{v}}. \quad (6.13')$$

Degenerate gas. The case of a degenerate gas is different. It may be seen from Figure 6.3(a) that for a degenerate gas, all quantum states to the left of ν_F are occupied by electrons. Because of that, the external field can act only on the electrons close to the Fermi level, lifting them to higher vacant levels by moving them from the left-hand region of the distribution to the right-hand region, as shown by the arrow 11'. This means that in a degenerate gas, only the electrons in the immediate vicinity of the Fermi level can take part in electrical conductivity. Therefore, one should take for the relaxation time in expressions (6.5) and (6.13), the relaxation time of the electrons whose energy is practically equal to the Fermi energy. Let us denote it by τ_F .

Substituting τ_F for τ in (6.5) and (6.13), we obtain the following expressions for

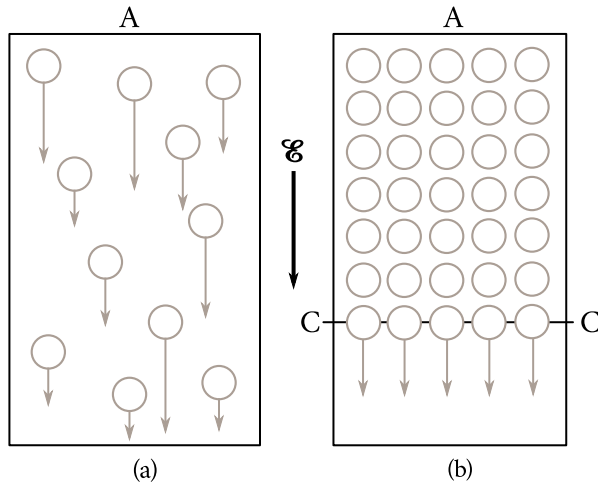


Figure 6.4: Mechanical model of the behaviour in of (a) a nondegenerate and (b) a degenerate electron gas; arrows denote the drift velocity in the electric held.

the electron mobility and the specific conductance of a degenerate gas:

$$u = \frac{q\tau_F}{m_n} = \frac{q\lambda_F v_F}{v_F}, \quad (6.5'')$$

$$\sigma = \frac{nq^2}{m_n} \tau_F = \frac{nq^2}{m_n} \frac{q\lambda_F v_F}{v_F}. \quad (6.13'')$$

where λ_F is the mean free path of the electrons with Fermi energy, v_F their velocity, and ν_F the number of collisions after which the directional velocity of such electrons becomes randomized.

Here is, a rough mechanical analogy to explain the different behaviour of a nondegenerate and a degenerate electron gas in an electric held.

Imagine small charged balls (“particles”) floating on the surface of water in a flat horizontal vessel A and moving at random with different velocities in the absence of an external field [Figure 6.4(a)]. Now, let us place this vessel in an external field \mathcal{E} . The resulting effect of the field on the ensemble of the balls, as a whole, will substantially depend on how closely the balls are packed on the surface of water. If the number of balls is small, so that the distances between them are large, every one of them will move freely, practically speaking, and will not interfere with the motion of its neighbours [Figure 6.4(a)]. In this instance the motion of the ensemble, as a whole, shall be determined by the average parameters of motion of the individual “particles”: by the average velocity \bar{v} , by the average elaxation time $\bar{\tau}$, by the mean free path $\bar{\lambda}$, etc.

If the balls are packed as closely as possible, so that there is no place for any

more balls on the water's surface, then the motion of the ensemble as a whole under the action of the field \mathcal{E} will be determined by the motion of the lower layer of the “particles”, CC, which separates the “occupied” states from the “vacant” states [Figure 6.4(b)], namely, by the velocity of those particles v_C , by the relaxation time t_C , by the mean free path λ_C , etc. In a degenerate electron gas, the part of this layer is played by electrons close to the Fermi level, which separates occupied states from the vacant ones.

§ 52. Wiedemann-Franz-Lorenz law

The transport of electric charge in an electric field is not the only result of the presence of the electron gas in a solid—another is heat transport in the presence of a temperature gradient. For this reason, it would be natural to expect that the electric and the heat conductivities of a solid are interrelated. This interrelation was first experimentally established by G. Wiedemann and P. Franz and theoretically explained by L. Lorenz for the case of metals. They showed the ratio of the heat conductivity \mathcal{K} of a metal to its specific conductance σ to be proportional to the absolute temperature T :

$$\frac{\mathcal{K}}{\sigma} = LT. \quad (6.14)$$

Expression (6.14) is the essence of the *Wiedemann-Franz-Lorenz law*; the proportionality factor L is called the *Lorenz number*.

The Wiedemann-Franz-Lorenz law can easily be obtained if one makes use of the expressions for \mathcal{K} and σ derived in the electron theory of metals. Dividing Eq. (4.58) for the heat conductivity of a metal (which is practically equal to its electron component) by Eq. (6.13''), we obtain

$$\frac{\mathcal{K}}{\sigma} = \frac{\pi^2}{3} \left(\frac{k_B}{q} \right)^2 T. \quad (6.15)$$

Comparing Eq. (6.15) with Eq. (6.14), we find that theoretical value of the Lorenz number

$$L = \frac{\pi^2}{3} \left(\frac{k_B}{q} \right)^2 = 2.45 \times 10^{-8} \text{ W } \Omega \text{ K}^{-2}. \quad (6.16)$$

Table 6.1 shows the experimental values of L for some pure metals at 0 °C. We see that the theoretical value of L agrees well with experiment.

In semiconductors with a nondegenerate electron gas the heat conductivity is not entirely due to the electrons. A substantial part of it is usually due to lattice conductivity. However, in this case too, the electron component of the semiconductor's heat conductivity obeys the Wiedemann-Franz-Lorenz law, the only dif-

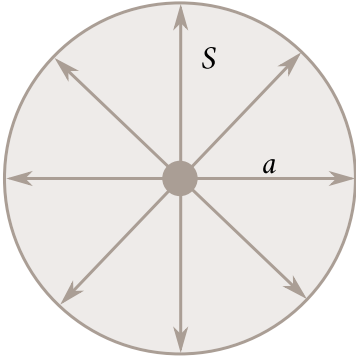


Figure 6.5: Thermal vibrations of lattice atoms (a is the amplitude of vibrations, and the shaded circle is effective scattering cross section).

ference being that its Lorenz number is not determined by Eq. (6.16) but is

$$L = 2 \left(\frac{k_B}{q} \right)^2. \tag{6.17}$$

§ 53. Temperature dependence of carrier mobility

Let us discuss now one of the principal problems of the theory of electrical conductivity of solids—the temperature dependence of carrier mobility. We shall discuss separately the high and the low temperature range.

High temperature range. In the high temperature range the dominant part is played by electron scattering on lattice vibrations phonons). Every lattice atom vibrates at random around its equilibrium position (Figure 6.5) remaining inside a sphere with a radius equal to the vibration amplitude a . The cross section of this sphere $S = \pi a^2$ may be taken as the scattering cross section of a vibrating atom: an electron moving in a conductor can run into one of such disks and be scattered.

Other conditions being equal, the probability for an electron to run into such a disk will, evidently, be proportional to its cross section, and the mean free path

Table 6.1

	Ag	Au	Cd	Cu	Ir	Mo	Pb
$L (10^8 \text{ W } \Omega \text{ K}^{-2})$	2.31	2.35	2.42	2.23	2.49	2.61	2.47

of the electron will be inversely proportional to that cross section:

$$\lambda \propto \frac{1}{S} \sim \frac{1}{a^2}.$$

The energy of a vibrating atom is proportional to the square of the amplitude: $E \propto a^2$. On the other hand, the average energy of atoms vibrating in a solid is proportional to the absolute temperature of the solid, T , that is, $E \propto T$. Therefore, in the high temperature range the mean free path of the electrons due to the thermal lattice vibrations should be inversely proportional to the absolute temperature of the body:

$$\lambda \propto \frac{1}{T}. \quad (6.18)$$

This result could have been obtained immediately with the aid of Eq. (4.37). According to this formula, the phonon concentration in a conductor in the high temperature range is proportional to T , that is, $n_{\text{ph}} \propto T$. For electron-phonon scattering, the electron mean free path should evidently be proportional to the phonon concentration and, consequently, inversely proportional to the absolute temperature T , that is, $\lambda \propto 1/n_{\text{ph}} \propto 1/T$. On the other hand, the average momentum of a phonon at high temperatures is so great that a single collision of the electron with a phonon (that is, at $\nu \approx 1$) already results in a practically total loss of the electron's initial velocity.

Substituting Eq. (6.18) into Eqs. (6.5') and (6.5'') and setting $\nu = 1$, we obtain the following expressions for the electron mobility:

$$u \propto \frac{\bar{\lambda}}{\bar{v}} \propto \frac{T^{-1}}{T^{1/2}} = T^{3/2}, \quad (\text{nondegenerate gas}) \quad (6.19)$$

$$u \propto \frac{\lambda_F}{v_F} \propto \frac{T^{-1}}{\text{constant}} \propto T^{-1}. \quad (\text{degenerate gas}) \quad (6.20)$$

Hence, in the high temperature range, where the dominant effect is scattering by phonons (by the lattice vibrations), the carrier mobility (of electrons or holes) in a nondegenerate gas is inversely proportional to $T^{3/2}$ and in a degenerate gas to T^{-1} . We see that in this instance too, the difference in the behaviour of the nondegenerate and the degenerate gases makes itself felt.

Low temperature range. In the low temperature range, the dominant effect is scattering by ionized impurity atoms. The mechanism of the scattering process is such that the impurity ions deflect electrons passing close to them and, thus, reduce their velocities in the original directions. As is shown in Figure 6.6, the velocity of the electron in the direction of the field was \mathbf{v}_0 before it was deflected by the positively charged ion. After the deflection it fell to \mathbf{v}'_0 .

The problem of charged centres deflecting charged particles was first solved

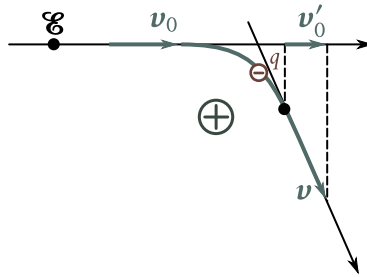


Figure 6.6: Electron q scattered by an ionized impurity atom (v_0 is electron velocity before scattering, and v after scattering).

by E. Rutherford who investigated the scattering of α -particles by the nuclei of chemical elements. Applied to our case, the formula for v obtained by Rutherford assumes the following form:

$$v \propto v^4 \left(\frac{\varepsilon}{Zq} \right)^2 m_n \quad (6.21)$$

where v is the electron velocity, ε the dielectric constant of the crystal, and Zq the charge of the scattering ion.

This result is quite understandable from qualitative considerations. The higher the electron velocity, their effective mass m_n and the field intensity reduction factor in the crystal (the greater ε is), the less the electrons will be deflected from their original path and the greater will be the number of collisions needed to randomize electron motion. Evidently, v should decrease with the increase in the charge of the scattering ion.

On the other hand, the mean free path of electrons being scattered by ionized impurity atoms is inversely proportional to their concentration and independent of temperature.

Taking this into account and substituting v from Eq. (6.21) into Eqs. (6.5') and (6.5''), we obtain the following:

$$u \propto \frac{\bar{v}\lambda}{\bar{v}} \propto \bar{v}^3 \propto T^{3/2}, \quad (\text{nondegenerate gas}) \quad (6.22)$$

$$u \propto \frac{v_F \lambda_F}{v_F} \propto v_F^3 = \text{constant}. \quad (\text{degenerate gas}) \quad (6.23)$$

Hence, in the low temperature range, the carrier mobility due to scattering by ionized impurity atoms is proportional to $T^{3/2}$ for conductors with a nondegenerate gas, and independent of T for conductors with a degenerate gas.

Figure 6.7(a) shows the temperature dependence of u for a nondegenerate gas, and Figure 6.7(b) an experimental $u(T)$ curve for silicon. It follows from Figure

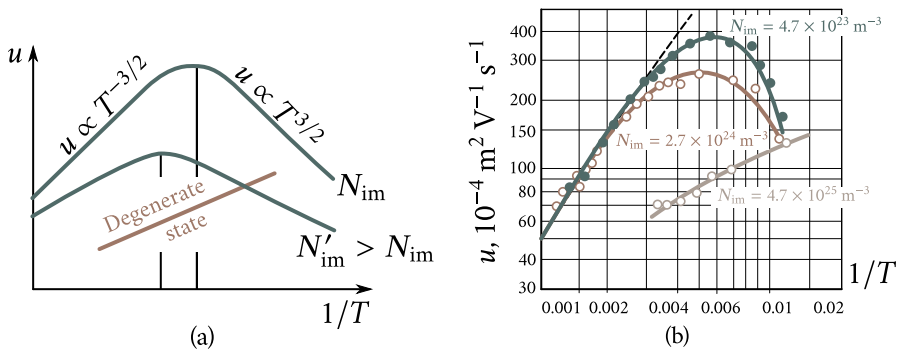


Figure 6.7: Temperature dependence of carrier mobility in semiconductors: (a)—theoretical curves; (b)—experimental curves for silicon doped with different amounts of phosphorus.

6.7 that experiment, on the whole, supports the conclusions of the theory as to the nature of the temperature dependence of carrier mobility in nondegenerate conductors.

We have discussed the case when in the low temperature range the main effect is due to scattering by ionized impurity atoms. However, for very pure and very perfect metals, containing negligible amounts of impurities and lattice imperfections, phonon scattering may turn out to be the principal charge-carrier scattering mechanism in the low temperature range. Let us find the temperature dependence of u for this case.

For electron-phonon scattering, the electron mean free path λ is inversely proportional to the phonon concentration n_{ph} . Since in the low temperature range $n_{\text{ph}} \propto T^3$ according to Eq. (4.36),

$$\lambda \propto \frac{1}{n_{\text{ph}}} \propto T^{-3}. \quad (6.24)$$

Now let us determine ν —the average number of collisions the electron should take part in to lose its original directional velocity.

At high temperatures, at which the average phonon momentum p_{ph} is equal, in order of magnitude, to the electron momentum p_e , $\nu \approx 1$. At low temperatures, $p_{\text{ph}} \ll p_e$, and as a result ν can be much greater than unity, being substantially dependent on temperature since p_{ph} rises with the rise in T .

Figure 6.8 shows the variation of the momentum of an electron that took part in an elastic collision with a phonon. The collision took place at point A. Before the collision, the electron's momentum was \mathbf{p}_e^0 and after the collision it became \mathbf{p}_e . Since the collision was an elastic one, the absolute value of the momentum has not

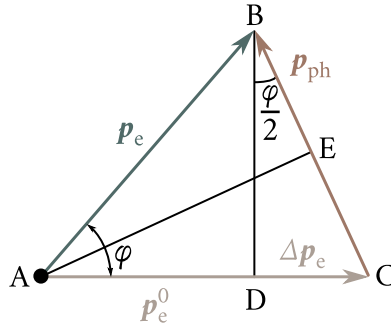


Figure 6.8: Calculating the number of collisions needed to nullify the electron's momentum in a given direction.

changed: $p_e^0 = p_e$. Only the direction has changed so that

$$\mathbf{p}_e = \mathbf{p}_e^0 + \mathbf{p}_{ph}.$$

The change in the direction of the electron's momentum brought about by the collision entails a reduction in its value in the original direction by Δp_e (Figure 6.8). It follows from $\triangle BCD$ that

$$\Delta p_e = p_{ph} \sin\left(\frac{\varphi}{2}\right)$$

where φ is the electron scattering angle. From $\triangle AEC$, it follows that $\sin(\varphi/2) = p_{ph}/(2p_e)$. Therefore,

$$\Delta p_e = \frac{p_{ph}^2}{2p_e}.$$

It is by this quantity that the electron momentum, in the original direction, is reduced as a result of a single collision with a phonon. To eliminate the electron momentum in the original direction altogether, the following number of collisions is needed:

$$\nu \approx \frac{p_e}{\Delta p_e} \approx 2 \left(\frac{p_e}{p_{ph}} \right)^2 \propto \frac{1}{p_{ph}^2}.$$

In the low temperature range, the energy of thermal lattice vibrations (the phonon gas energy) is, according to Eqs. (4.30) and (4.36), $E_{\text{lattice}} \propto T^4$ and the phonon gas concentration is $n_{ph} \propto T^3$. Therefore, the average phonon energy

$$\bar{\varepsilon}_{ph} = \frac{E_{\text{lattice}}}{n_{ph}} \propto T$$

increases in proportion to T . Since the phonon momentum is

$$p_{ph} = \frac{\bar{\varepsilon}_{ph}}{v}$$

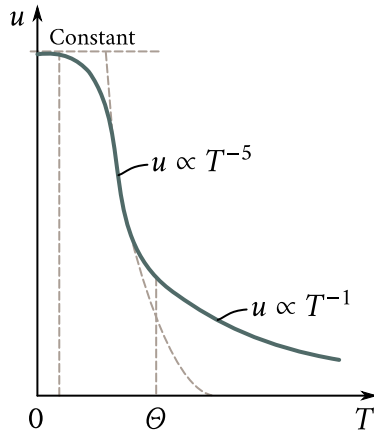


Figure 6.9: Temperature dependence of free electron mobility in pure metals.

(v is the velocity of sound in the crystal), the phonon momentum in this temperature range is also proportional to T :

$$p_{\text{ph}} \propto T. \quad (6.25)$$

Therefore,

$$\nu \propto \frac{1}{p_{\text{ph}}^2} \propto T^{-2}. \quad (6.26)$$

Substituting λ , from Eq. (6.24) and ν from Eq. (6.26) into Eq. (6.5''), we obtain the following expression for free carrier mobility in pure metals in the low temperature range:

$$u \propto \frac{\nu_F \lambda_F}{\nu_F} \propto T^{-5}. \quad (6.27)$$

Figure 6.9 shows the qualitative curve of the temperature dependence of u for pure metals. In the high temperature range (above the Debye temperature Θ) the carrier mobility $u \propto T^{-1}$, in the low temperature range (much below Θ) $u \propto T^{-5}$. In the intermediate temperature range, a gradual transition from the T^{-1} to the T^{-5} dependence takes place. Finally, close to absolute zero, the thermal vibrations become so weak that carrier scattering by impurity atoms and lattice defects, which are always present in a metal no matter how pure and perfect it is, becomes of primary importance. In this case, the carrier mobility ceases to depend on temperature [see Eq. (6.23)] and the u versus T curve follows a line parallel to the temperature axis.

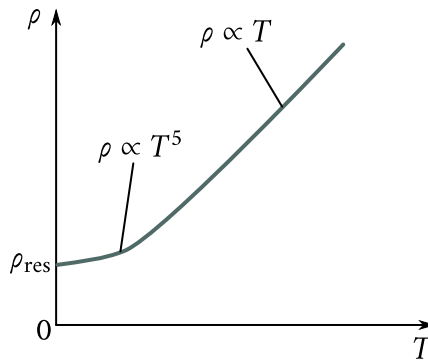


Figure 6.10: Schematic plot of the temperature dependence of specific resistance of pure metals.

§ 54. Electrical conductivity of pure metals

Electrical conductivity of pure metals is due to the drift of free charge carriers of one sign. In the absolute majority of pure metals, the charge carriers are free electrons. However, in some metals, such as beryllium and zinc, the charge is carried by holes.

The conductivity, or specific conductance, of electron metals is described by Eq. (6.12):

$$\sigma = qnu.$$

Since the metals are degenerate conductors, the electron concentration in them is practically independent of temperature. Because of that, temperature dependence of specific conductance is determined entirely by the temperature dependence of the mobility of the electrons in a degenerate electron gas, as discussed in the previous section.

Substituting u from Eqs. (6.20) and (6.27) into (6.12), we obtain the following expression for σ and the specific resistance ρ of pure metals:

$$\sigma = \frac{A}{T}, \quad \rho = aT, \quad (\text{high temperature range}) \quad (6.28)$$

$$\sigma = \frac{B}{T^5}, \quad \rho = bT^5. \quad (\text{low temperature range}) \quad (6.29)$$

Here, A , B , a , and b are proportionality factors.

Figure 6.10 shows schematically the dependence of specific resistance of pure metals on temperature. In the high temperature range, this dependence is represented by a straight line, in the low temperature range, by a parabola of the fifth degree, and in the vicinity of absolute zero, by a straight line parallel to the tem-

perature axis.

A more rigorous quantum mechanical calculation enables the coefficients A , B , a and b in formulae (6.28) and (6.29) to be found. Table 6.2 shows the specific conductance of some pure metals at room temperature, calculated (σ_{theory}) and measured experimentally ($\sigma_{\text{experiment}}$) (in units of $10^6 \Omega^{-1} \text{ m}^{-1}$).

It follows that for Na and K, in which the conducting electrons are almost in a free state, the agreement between theory and experiment is satisfactory. As the atomic mass increases, so does the lattice potential and the interaction of the conducting electrons with the lattice. This means that the free electron approximation becomes less valid. The result is a discrepancy between σ_{theory} and $\sigma_{\text{experiment}}$ which grows.

Table 6.3 shows the ratios of the specific conductivity of gold σ_0 at 273 K to σ at low temperatures, calculated and measured.

The agreement between theory and experiment seems to be quite satisfactory.

§ 55. Electrical conductivity of metal alloys

In metal alloys too, the carrier concentration is independent of temperature. Therefore, the temperature dependence of specific conductance in alloys is determined entirely by the temperature dependence of the carrier mobility. Let us discuss this problem in more detail.

Suppose that some sites of an ideal metal lattice, for instance, of a copper lattice, with a strictly periodic potential [Figure 6.11(a)] are at random replaced by atoms

Table 6.2

	Na	K	Rb	Cu	Ag	Au
σ_{theory}	22	19	20	100	90	107
$\sigma_{\text{experiment}}$	23	15	8	64	67	68

Table 6.3

	273 K	87.4 K	57.8 K	20.4 K	11.1 K	4.2 K
$(\sigma_0/\sigma)_{\text{theory}}$	1	0.2645	0.1356	0.0060	0.0003	3×10^{-6}
$(\sigma_0/\sigma)_{\text{experiment}}$	1	0.2551	0.1314	0.0058	0.0003	3×10^{-6}

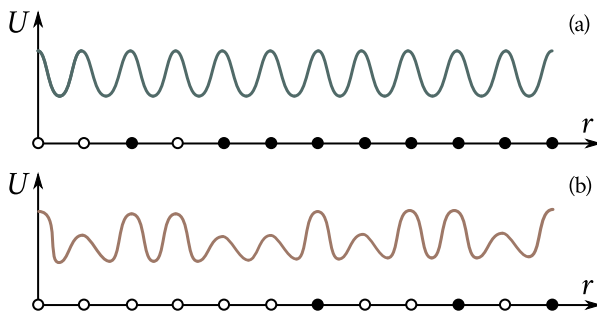


Figure 6.11: Violation of lattice potential's periodicity by impurity atoms: (a)—strictly periodic potential of ideal lattice built of atoms of one kind; (b)—violation of potential's periodicity by impurity atoms substituting matrix atoms at random.

of some other element, say, gold. Since the potential of the field of an impurity atom is not the same as that of the matrix atom, the lattice potential will cease to be strictly periodic [Figure 6.11(b)]. It will be distorted by the disordered impurity atoms. Naturally, such distortion will lead to carrier scattering and to the appearance of additional electrical resistance.

It was L. Nordheim who demonstrated that, in the simplest case of binary alloys of the solid solution type, the carrier mobility due to scattering on lattice imperfections is described by the following approximate relation:

$$u_{\text{al}} \propto \omega(1 - \omega) \quad (6.30)$$

where ω and $(1 - \omega)$ are the fractional parts of the metals constituting the alloy.

Substituting u_{al} from Eq. (6.30) into Eq. (6.12) and keeping in mind that $\rho = 1/\sigma$, we obtain the following expression for the specific resistance of a binary alloy:

$$\rho_{\text{al}} = \beta[\omega(1 - \omega)] \quad (6.31)$$

where β is a proportionality factor.

The function $\omega(1 - \omega)$ has a maximum at $\omega = 1/2$, that is when the concentrations of both components are equal. Figure 6.12(a) shows the dependence of the specific resistance of copper-gold alloys on the gold contents. The curve passes through a maximum corresponding to a 50% contents of copper (or gold) in the alloy.

It also follows from Figure 6.12(a) that ρ_{al} is much greater than that of the pure components. For instance, at room temperature, $\rho_{\text{Cu}} = 1.7 \times 10^{-8} \Omega \text{ m}$ and $\rho_{\text{Au}} = 1.56 \times 10^{-8} \Omega \text{ m}$, whereas $\rho_{50\% \text{Cu} + 50\% \text{Au}} = 15 \times 10^{-8} \Omega \text{ m}$. This is quite natural since the disorder in the lattice has a much more detrimental effect on the lattice periodicity than the thermal vibrations. If, however, the alloyed met-

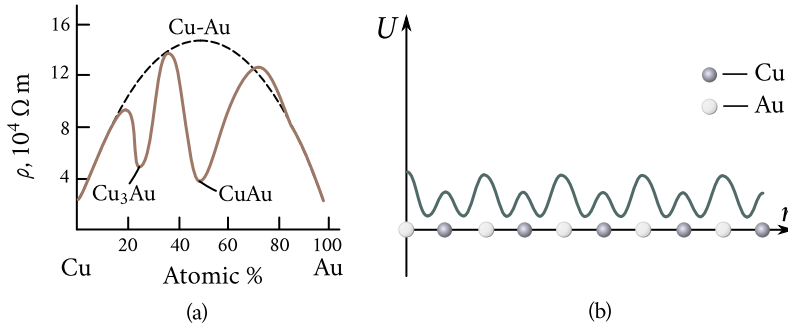


Figure 6.12: (a)—dependence of specific resistance of solid solutions of gold and copper on composition, (b)—lattice potential's periodicity recovered in the ordering of the structure.

als taken in appropriate proportions form ordered alloys, or metallic compounds, with an ordered structure, the lattice periodicity is recovered [Figure 6.12(b)] and the resistance due to impurity scattering vanishes practically altogether. For the copper-gold alloys, the appropriate concentrations are those which correspond to the stoichiometric composition of Cu_3Au and CuAu [Figure 6.12(a), solid curves]. This may serve as a proof of the validity of the quantum theory of electrical conductivity, which maintains that the cause of the electrical resistance of solids is not the collision of the free electrons with the lattice atoms but their scattering by the lattice defects which distort the periodic lattice potential. An ideal regular imperfection-free lattice with a strictly periodic potential is incapable of scattering free charge carriers and must therefore have zero resistance. This conclusion is supported by numerous experiments carried out with extremely pure metals in the low temperature range, the relevant data being presented in Table 6.3: as the degree of purity of a metal is increased its specific resistance near absolute zero diminishes, continuously tending to zero. We would like to stress that, this is not the phenomenon of superconductivity which we shall discuss later, but the natural behaviour of absolutely all pure metals in the extreme low temperature range, which is a consequence of the quantum mechanical nature of electrical resistance.

For small impurity contents one may set in Eq. (6.31) $(1 - \omega) \approx 1$. Then, $\rho_{\text{al}} \propto \omega$. This specific resistance is independent of temperature and does not vanish at absolute zero. It is termed *residual resistivity* ρ_{res} (see Figure 6.10).

At temperatures other than absolute zero, a resistivity ρ_{T} due to electron scattering by the lattice vibrations is added to the residual resistivity and the total resistivity becomes

$$\rho = \rho_{\text{res}} + \rho_{\text{T}}. \quad (6.32)$$

This relation expresses *Matthiessen's rule*, which speaks of the additivity of specific

resistance.

Let us discuss now the *temperature coefficient of resistivity* α . As is well known, it expresses the relative variation of the specific resistance of a conductor whose temperature is raised by 1 K. For pure metals, $\rho = \rho_T$ and therefore,

$$\alpha = \frac{1}{\rho_T} \frac{d\rho_T}{dT}. \quad (6.33)$$

Experiment shows α roughly to be

$$\alpha \approx \frac{1}{273} \text{ K}^{-1} \approx 0.004 \text{ K}^{-1}$$

[see Table 6.4]. For alloys, $\rho = \rho_{\text{res}} + \rho_T$; therefore,

$$\alpha_{\text{al}} = \frac{1}{\rho} \frac{d\rho}{dT} = \frac{1}{(\rho_{\text{res}} + \rho_T)} \frac{d\rho_T}{dT}$$

since ρ_{res} is independent of temperature. This expression may be transformed into

$$\alpha_{\text{al}} = \frac{1}{\left(1 + \frac{\rho_{\text{res}}}{\rho_T}\right)} \frac{1}{\rho_T} \frac{d\rho_T}{dT} = \frac{\alpha}{\left(1 + \frac{\rho_{\text{res}}}{\rho_T}\right)} \quad (6.34)$$

where α is the temperature coefficient of resistivity of pure metals.

It follows from Eq. (6.34) that α_{al} should be less than α of a pure metal, the less the greater ρ_{res} is in comparison with ρ_T . Usually, ρ_{res} is an order of magnitude or more greater than ρ_T . Therefore, α_{al} may be an order of magnitude or less than α of a pure metal, and this is on the whole supported by experiment (Table 6.4; the data are for room temperature).

However, in many cases, the temperature dependence of an alloy's resistance is much more complex than that which follows from the simple additive rule (6.32), and the temperature coefficient of resistivity of some alloys may be much less than one could expect from Eq. (6.34). More than that, it does not remain constant in a wide temperature interval but may in some cases even become negative as is, for instance, the case with constantan (Table 6.4) and with some other alloys.

Table 6.4

	Bronze				Nichrome	Constantan
	Cu	Sn	Ni	(88% Cu, 18% Sn, 1% Pb)	(80% Ni, 20% Cr)	(54% Cu, 46% Ni)
α (10^3 K^{-1})	4.1	4.2	6.2	0.5	0.13	-0.004

A high specific resistance together with a low temperature coefficient of resistivity made alloys valuable materials for the production of various wire and film resistors and variable resistors (rheostats) widely used in different fields of science and technology.

§ 56. Intrinsic conductivity of semiconductors

The electrical conductivity of very pure and perfect single crystal semiconductors, in the not very low temperature range, is due to intrinsic charge carriers, that is, to electrons and holes. Such conductivity is termed *intrinsic*.

Since there are two types of carriers in the intrinsic semiconductor, electrons and holes, its specific conductance is the sum of the conductivity $\sigma_n = qn_i u_n$ due to free electrons, with the concentration n_i and the mobility u_n , and of the conductivity $\sigma_p = qp_i u_p$ due to the presence of holes, with the concentration p_i and the mobility u_p . Since $n_i = p_i$, the total specific conductance of an intrinsic semiconductor is

$$\sigma_i = \sigma_n + \sigma_p = qn_i(u_n + u_p). \quad (6.35)$$

According to Eq. (5.37), the hole (or electron) concentration in an intrinsic semiconductor is

$$n_i = p_i = 2 \left(\frac{2\pi \sqrt{m_n m_p} k_B T}{h^2} \right)^{3/2} e^{-E_g/(2k_B T)}.$$

The carrier mobility in the intrinsic conductivity range is given by Eq. (6.19). Substituting Eqs. (5.37) and (6.19) into (6.35) we obtain

$$\sigma_i = \sigma_0 e^{-E_g/(2k_B T)} \quad (6.36)$$

where σ_0 denotes the preexponential expression.

It follows from Eq. (6.36) that, $\sigma_i \rightarrow \sigma_0$ as $T \rightarrow \infty$. We thus conclude that, if the rule (6.36) remains valid for infinitely high temperatures, σ_0 would be the specific conductance of the semiconductor as $T \rightarrow \infty$.

The temperature dependence of σ_i can be conveniently represented in the semilogarithmic coordinates. Taking the logarithm of Eq. (6.36), we obtain

$$\ln \sigma_i = \ln \sigma_0 - \left(\frac{E_g}{2k_B T} \right). \quad (6.37)$$

If we plot $1/T$ along the x axis and $\ln \sigma$ along the y axis, we will obtain a straight line that cuts off a section $\ln \sigma_0$ [Figure 6.13(a)] on the y axis. The tangent of α of this straight line to the x axis is equal to $E_g/(2k_B T)$. Plotting this dependence, we can find the constant σ_0 and the width of the forbidden band E_g . Figure 6.13(b) shows the experimental $\ln \sigma_i$ versus $1/T$ dependence for pure germanium and silicon. The

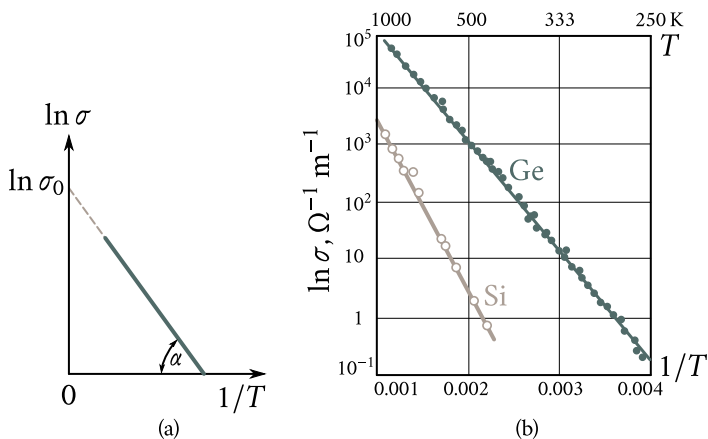


Figure 6.13: Temperature dependence of intrinsic conductivity of a semiconductor: (a)—theoretical curve; (b)—experimental plots for germanium and silicon.

forbidden bands as determined from the inclination angles of the curves turned out to be 0.72 eV and 1.2 eV wide.

Comparing the results of this section with those of the previous one, we see that there is the following principal difference between metals and semiconductors. In metals, where the electron gas is in a degenerate state, the carrier concentration is practically independent of temperature and the temperature dependence of conductivity is determined entirely by the temperature dependence of carrier mobility. In the semiconductors, on the other hand, the carrier gas is nondegenerate and its concentration depends strongly on temperature [see Eq. (5.37)]. Because of that, their conductivity is entirely determined by the temperature dependence of carrier concentration [see Eq. (6.36)].

For a specific temperature the carrier concentration and the conductivity of a semiconductor are determined by the width of its forbidden band. This may be seen quite clearly from the data of Table 6.5 which contains the widths of the forbidden bands and the specific resistances of the elements of Group IV of the Mendeleev periodic table, the elements having the diamond-type lattice. As the width of the forbidden band decreases from 5.2 eV (diamond) to 0.08 eV (gray tin), the room temperature specific resistance diminishes by 16 orders of magnitude.

§ 57. Impurity (extrinsic) conductivity of semiconductors

The temperature dependence of specific conductance of nondegenerate impurity semiconductors, as that of intrinsic semiconductors, is for the most part deter-

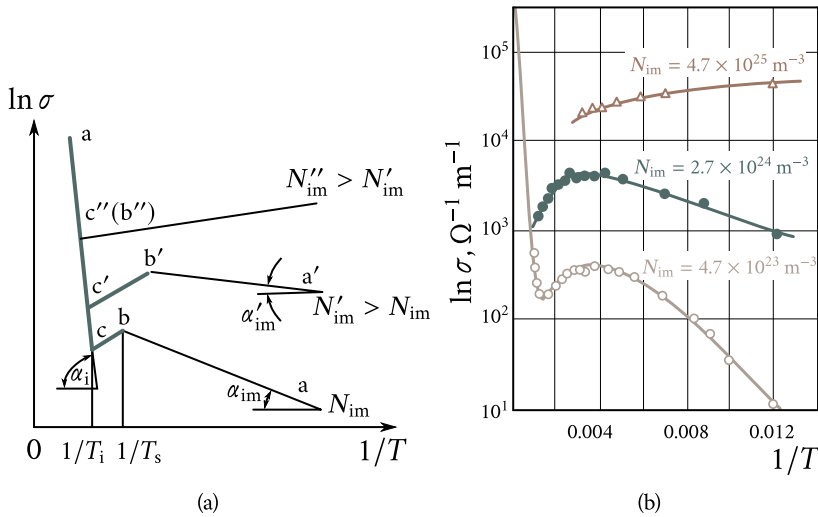


Figure 6.14: Temperature dependence of specific conductance of impurity: semiconductors: (a)—theoretical curve; (b)—experimental plots for silicon containing different amounts of phosphorus.

mined by the temperature dependence of carrier concentration. Because of this, the curve representing the temperature dependence of σ must at least qualitatively be analogous to the n versus T curve, where the latter is shown in Figure 5.22.

The temperature dependence of $\ln \sigma$ for an impurity semiconductor is represented qualitatively in Figure 6.14(a). There are three distinct regions on this curve ab , bc , and cd .

The region ab lies between absolute zero and the impurity saturation temperature T_s . The carrier concentration in this region is described by Eq. (5.41):

$$n = \sqrt{2N_d} \left(\frac{2\pi m_n k_B T}{h^2} \right)^{3/2} e^{-E_d/(2k_B T)}.$$

The mobility is determined mainly by impurity and imperfection scattering, and according to Eq. (6.22), is proportional to $T^{3/2}$. Substituting Eqs. (5.41) and (6.22)

Table 6.5

	Diamond	Silicon	Germanium	Gray tin
E_g (eV)	5.2	1.12	0.66	0.08
ρ (Ωm)	10^{10}	3×10^3	0.47	2×10^{-6}

into (6.12) we obtain

$$\sigma_{\text{im}} = \sigma_{\text{im}}^0 e^{-E_d/(2k_B T)} \quad (6.38)$$

where σ_{im}^0 is a factor that depends weakly on temperature (as compared with the exponential).

Taking the logarithm of Eq. (6.38), we obtain

$$\ln \sigma_{\text{im}} = \ln \sigma_{\text{im}}^0 - \left(\frac{E_d}{2k_B T} \right). \quad (6.39)$$

In the $\ln \sigma_{\text{im}}$ versus $1/T$ coordinate system we get a straight line which makes an angle α_{im} with the $1/T$ axis such that $\tan \alpha_{\text{im}} = E_d/(2k_B)$ is proportional to the impurity ionization energy E_d . Hence, region ab corresponds to impurity, or extrinsic, conductivity, which is due to impurity carriers freed as the result of the ionization of impurity atoms.

The region bc lies between the impurity saturation temperature T_a and the temperature of intrinsic conductivity T_i . In this range, all the impurity atoms are ionized but no noticeable excitation of intrinsic carriers takes place. Because of that, the carrier concentration remains approximately constant and equal to the impurity concentration $n \approx N_d$. Therefore, in this region the temperature dependence of the conductivity is determined by the temperature dependence of carrier mobility. If the principal carrier scattering mechanism inside this region is the scattering on thermal lattice vibrations, which causes the mobility to fall with temperature, then the specific conductance will also diminish with the rise in temperature. This is just the case shown in Figure 6.14(a). But if the principal mechanism is impurity or imperfection scattering, then the specific conductance in the region bc will increase with temperature.

The region cd corresponds to the transition to intrinsic conductivity. Inside this region, the carrier concentration is equal to the intrinsic carrier concentration. Therefore, the conductivity of the semiconductor in this region is

$$\sigma \approx \sigma_{\text{im}} = \sigma_{\text{im}}^0 e^{-E_d/(2k_B T)}.$$

In semilogarithmic coordinates $\ln \sigma$ versus $1/T$ this dependence is represented by a straight line cd, making an angle α_i with the $1/T$ axis, its tangent being proportional to the width of the forbidden band: $\tan \alpha_i = E_d/(2k_B)$.

Figure 6.14(b) shows the temperature dependence of the conductivity of phosphorus-doped silicon. A comparison with Figure 6.14(a) shows that in the simplest cases the theory ensures a qualitative agreement with experiment.

Thermistors. The strong dependence of the resistance of semiconductors on the temperature is utilized in a wide class of semiconductor devices, the thermistors. They are bulk semiconductor resistors with a large temperature coefficient

of resistivity and a nonlinear current-voltage characteristic.

Thermistors are used in measuring temperature and power of ultrahigh frequency radiation, for temperature compensation in various electric circuits, for timing relays, etc. Microthermistors, which have small dimensions and low thermal inertia, are being used in the study of heat exchange processes in plants and living organisms including early diagnosis of human illnesses. The use of a thin semiconducting film in a bolometer, made it possible to increase its sensitivity to 10^{-10} W. Such bolometers placed in the focus of a parabolic mirror are capable of detecting aircraft, tanks, ships and other bodies that radiate heat at a distance of the order of several kilometers. A highly sensitive semiconductor bolometer detected infrared radiation reflected by the moon's surface.

§ 58. Deviation from Ohm's law. The effect of a strong field

The proportionality between the current density \mathbf{i} and the field intensity \mathcal{E} demanded by the Ohm's law (6.1) remains as long as σ , which enters this law as a proportionality factor, remains independent of \mathcal{E} .

Let us find what are the conditions in which this requirement is fulfilled.

According to Eq. (6.5'), the carrier mobility in nondegenerate semiconductors $u \propto \lambda/v$, where v is the resultant velocity of carrier motion. It is the sum of the thermal v_0 and drift v_d velocities:

$$\mathbf{v} = \mathbf{v}_0 + \mathbf{v}_d.$$

For weak fields

$$v_d \ll v_0 \tag{6.40}$$

the resultant carrier velocity is v_0 and is independent of \mathcal{E} . Therefore, both the carrier mobility u and their concentration and, consequently, the specific conductance $\sigma = qnu$ are independent of \mathcal{E} . Such fields are termed *weak*.

Hence, Ohm's law, which requires a linear dependence of \mathbf{i} on \mathcal{E} , is valid only in the case of weak fields complying with the condition (6.40).

As the field \mathcal{E} increases, the drift velocity v_d rises, and in fields of high intensity, may become comparable in order of magnitude with v_0 . In this case, the resultant velocity begins to be dependent on \mathcal{E} and because of that, the mobility u and the specific conductance σ too become dependent on \mathcal{E} . Naturally, the result is a distortion of the linear dependence of \mathbf{i} on \mathcal{E} , that is, a deviation from Ohm's law. Fields in which such phenomena take place are termed *strong*.

Calculations show that when scattering by thermal lattice vibrations is the

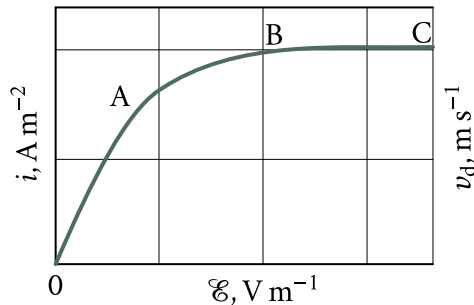


Figure 6.15: Nonlinear carrier drift velocity in semiconductors: $i \propto \mathcal{E}$ (Ohm's law) in region OA, $i \propto \sqrt{\mathcal{E}}$ in region AB, and i is independent of \mathcal{E} in region BC (saturation of drift velocity).

principal scattering mechanism in strong fields

$$u \propto \mathcal{E}^{-1/2}, \quad \sigma = qnu \propto \mathcal{E}^{-1/2}, \quad i = \sigma \mathcal{E} \propto \mathcal{E}^{1/2}. \quad (6.41)$$

In still stronger fields, the drift velocity v_d ceases to be dependent on the so-called *drift velocity saturation effect* sets in (Figure 6.15). Since $i \propto v_d$, such fields are also characterized by current saturation. The current-voltage characteristic of a semiconductor becomes distinctly nonlinear.

The rise in the resultant electron velocity in an external field is equivalent to the rise in the temperature of the electron gas. Therefore, this effect is known as electron gas heating and the electrons whose average kinetic energy exceeds that of the lattice atoms are termed *hot electrons*.

Strong fields can bring about not only changes in carrier mobility but in their concentration as well. There are several mechanisms leading to that result.

Thermoelectron ionization. In strong fields, not only free electrons become heated but to a lesser extent, bound electrons too. Therefore, the probability of their transition to the conduction band increases in the same way as it would increase if the temperature of the semiconductor as a whole would be raised by an appropriate amount. This results in an increase in free carrier concentration and in the specific conductance of the semiconductor, σ . Such phenomenon became known by the name of *thermoelectron ionization*. Its theory was developed by Ya. I. Frenkel.

Impact ionization. If conduction electrons of a heated electron gas receive enough energy to ionize neutral atoms lifting their electrons to the conduction band and themselves remaining in the conduction band, then there will be an avalanche-type increase in the free carrier concentration, until the process is counterbalanced by recombination. This mechanism of free carrier breeding is termed *impact ionization*.

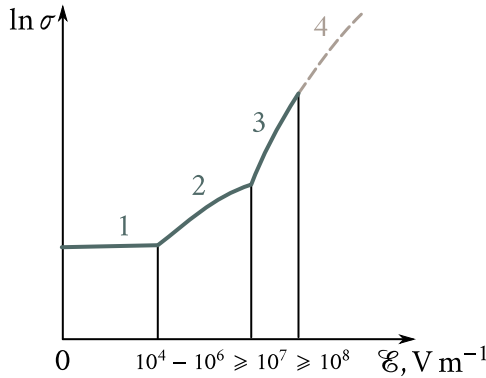


Figure 6.16: Qualitative dependence of specific conductance of germanium on electric field intensity: 1—ohmic region, 2—Frenkel region, 3—electrostatic ionization region, 4—breakdown region.

Electrostatic ionization. In high-intensity fields the transition of the electrons from the valence band to the conduction band by means of tunnelling through the forbidden band becomes possible. This effect is known as the *Zener effect*, or *electrostatic ionization*. The probability of tunnelling and, consequently, the tunnel current density increase drastically with the increase in the field intensity and decrease with the increase in the width of the forbidden band.

Figure 6.16 shows a qualitative curve of the variation of specific conductance of germanium with field intensity (in semilogarithmic approximate coordinates). Also shown are limits within which those mechanisms of carrier generation resulting in the increase in electric conductivity operate (1, 2—the ohmic and the Frenkel regions; 3, 4—the regions of electrostatic ionization and breakdown).

§ 59. The Gunn effect

It was demonstrated in the previous section that in strong fields there is a phenomenon of nonlinear drift velocity: the drift velocity changes not in direct proportion to the field intensity \mathcal{E} , the result being a deviation from Ohm's law.

An interesting effect of nonlinear drift velocity in gallium arsenide was discovered by J. B. Gunn. It became known as the *Gunn effect*. Figure 6.17(a) shows the pattern of the conduction band of gallium arsenide. It has two minimums in the $[100]$ direction: one at $k = 0$ and the other at $k = 0.8k_0$ (k_0 is the wave vector corresponding to the Brillouin zone boundary). The second minimum is $E = 0.36$ eV above the first. In normal conditions the electrons of the conduction band occupy the first minimum, where their effective mass is $m'_n = 0.072m$ and the mobility

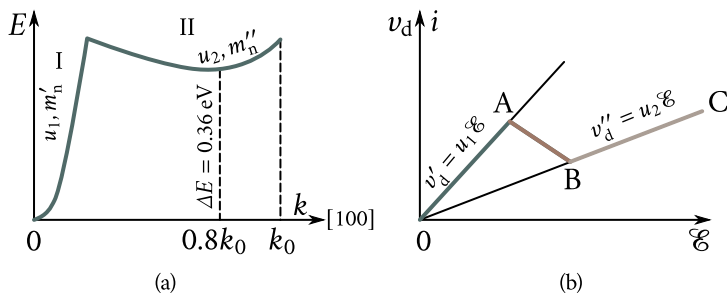


Figure 6.17: The Gunn effect: (a)—structure of conduction band in allium arsenide in $[100]$ direction; (b)—variation of drift velocity and current density with the increase in electric field intensity.

$u_1 = 0.5 \text{ V m}^{-2} \text{ s}^{-1}$. When an external field is applied to the crystal, the electrons' drift velocity becomes $v'_d = u_1 \mathcal{E}$, which increases in proportion to \mathcal{E} [the straight line OA, Figure 6.17(b)]. This goes on until the heated electrons accumulate sufficient energy to go over to the upper minimum, where their effective mass is much greater ($m''_n = 1.2m$) and the mobility much lower ($u_2 = 0.01 \text{ V m}^{-2} \text{ s}^{-1}$). Such a transition results in a drastic reduction in the drift velocity (because of a lower electron mobility) and the current density, that is, in the appearance of the region AB with a negative differential conductivity ($\sigma_{\text{dif}} = di/dV$). After most of the electrons move over to the upper minimum any further, increase in \mathcal{E} will be accompanied by an increase in drift velocity $v''_d = u_2 \mathcal{E}$ and by a proportional increase in the current density i (region BC).

The presence of a negative differential conductivity region on the current-voltage characteristic of a gallium arsenide crystal makes it possible to devise, on the basis of the Gunn effect, ultra-high frequency (UHF) oscillators known as *Gunn diodes*.

The Gunn effect was first discovered in 1963. In 1966, a first commercial type of an UHF generator working at a frequency of 2 GHz to 3 GHz with a power output of approximately 100 W in pulsed operation, was produced. At an electronic instrumentation and automatics exhibition which took place in the United States in 1968, radars using Gunn generators to measure the speed of moving objects were displayed. Those radars were so small that they could be carried by hand.

§ 60. Photoconductivity of semiconductors

Let us turn a ray of light of intensity I_0 on the semiconductor [Figure 6.18(a)]. Passing through the semiconductor the light is gradually absorbed and its intensity is

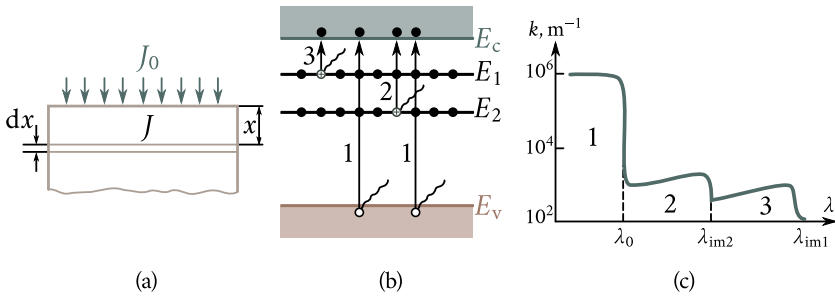


Figure 6.18: Generation of free charge carriers by light (internal photoeffect): (a)—absorption of light by a semiconducting specimen; (b)—excitation of free charge carriers from valence band, 1 (intrinsic absorption), and from impurity levels, 2 and 3 (impurity absorption); (c)—dependence of absorption coefficient on wavelength (1—intrinsic absorption band, 2 and 3—impurity absorption bands, λ_0 —threshold of photoeffect, λ_{im1} and λ_{im2} —thresholds of impurity photoconductivity).

diminished. Cut out an infinitely thin layer dx at a distance x from the semiconductor's surface. The amount of luminous energy dJ absorbed in the layer dx is proportional to the intensity J of the light passing through this layer and to its thickness dx :

$$dJ = -kJ dx. \quad (6.42)$$

The minus sign shows that the energy diminishes; the term for the proportionality factor k is *absorption coefficient*. For $dx = 1$, $k = -dJ/J$. Thus, the absorption coefficient is numerically equal to the relative variation of the intensity of light passing through an absorbing medium of unit thickness. Its dimensions are reciprocal to length (m^{-1}).

Integrating Eq. (6.42), we obtain

$$J = J_0 e^{-kx}. \quad (6.43)$$

The light absorbed in a semiconductor may be the cause of generation of excess carriers, which increase the total free carrier concentration. The arrows 1 in Figure 6.18(b) show the process of excitation of the conduction electrons and holes in the course of intrinsic absorption of light by a semiconductor. A photon with an energy $h\nu$ equal or greater than the forbidden band width E_g , transports an electron from the valence band into the conduction band. The generated electron-hole pairs are free and can take part in the semiconductor's conductivity.

To excite electrons from the levels of impurity atoms the photon energy should be $h\nu \geq E_{im}$, where E_{im} is the ionization energy of those atoms. Such impurity levels in Figure 6.18(b) are E_1 and E_2 ; the process of electron excitation from these levels is shown by arrows 2 and 3.

Thus, if

$$\begin{aligned} h\nu &\geq E_g \text{ in case of intrinsic semiconductors and} \\ h\nu &\geq E_{\text{im}} \text{ in case of impurity semiconductors,} \end{aligned} \quad (6.44)$$

then, excess charge carriers are generated in the semiconductor and its conductivity increases.

The process of internal liberation of electrons due to the action of light is termed the *internal photoeffect*. The additional conductivity of a semiconductor irradiated with light is termed *photoconductivity*. The name for initial conductivity due to the thermal carrier excitation is termed *dark conductivity*, for it is the conductivity of the semiconductor kept in darkness. Light can excite excess carriers both, from the intrinsic and from the impurity levels, and accordingly, two types of conductivity can be distinguished: the *intrinsic* and the *impurity*. Using Eq. (6.44), we can find the threshold of this process, that is, the maximum wavelength of light that is still photoelectrically active:

$$\begin{aligned} \lambda_0 &= \frac{ch}{E_g} \text{ for the intrinsic semiconductors and} \\ \lambda_{\text{im}} &= \frac{ch}{E_{\text{im}}} \text{ for impurity semiconductors,} \end{aligned} \quad (6.45)$$

where c is the velocity of light.

The ionization energy for photoconductivity in pure semiconductors E_g lies in the range of 0.1 eV to 5 eV, the majority having $E_g \approx 1$ eV to 3 eV. The threshold for the latter lies in the visible part of the spectrum. Many impurity semiconductors have E_{im} of the order of decimal fractions of an electron volt and even lower. For them the threshold lies in the infrared part of the spectrum.

Figure 6.18(c) shows a schematic dependence of the absorption coefficient k on the wavelength λ for a semiconductor with two impurity levels E_1 and E_2 [Figure 6.18(b)]. The absorption spectrum of such a semiconductor consists of three absorption bands: the intrinsic absorption band 1 corresponding to the electron transition from the valence to the conduction band, and two impurity bands (2 and 3). They correspond to electron transitions from the impurity levels E_1 and E_2 to the conduction band [Figure 6.18(b)]. Light with $\lambda < \lambda_0 = hc/E_g$ is practically completely absorbed near the surface in a layer $x \approx 10^{-6}$ m thick; its absorption coefficient is $k \approx 10^6 \text{ m}^{-1}$. The impurity absorption coefficient depends on the concentration of impurities but seldom exceeds $k \approx 10^3 \text{ m}^{-1}$. The less the impurity ionization energy E_g the greater the maximum wavelength of photoconductivity is according to Eq. (6.45).

The impurity photoeffect is only possible if the impurity levels E_1 and E_2 are occupied by electrons, that is, if the semiconductor's temperature is below the

temperature of impurity exhaustion, T_s . For this reason, one usually has to cool the semiconductor to be able to observe photoconductivity, the necessary cooling temperatures being the lower the greater the maximum wavelength. For instance, gold-doped germanium has $\lambda_{\text{im}} = 9 \mu\text{m}$ and must have liquid nitrogen cooling ($T = 78 \text{ K}$); germanium doped with the elements of Groups III or V of the Mendeleev periodic table has $\lambda_{\text{im}} = 100 \mu\text{m}$ and needs liquid helium cooling ($T = 4.2 \text{ K}$).

If the intensity of light entering the semiconductor is J , the amount of luminous energy (the number of photons) absorbed in a unit volume of the semiconductor per unit time will be kJ , and the rate of carrier generation will be

$$g = Jk\beta \quad (6.46)$$

where β is the quantum yield, which shows the number of free carriers generated by an absorbed photon.

In the absence of recombination, the number of excess carriers would grow continuously with time. The effect of recombination, whose rate rises with the concentration of excess carriers, is to establish a stationary state in the semiconductor when the generation rate is equal to the recombination rate [see Eq. (5.49)]:

$$g = R = \frac{\Delta n_0}{\tau}. \quad (6.47)$$

This state is characterized by a constant (stationary) excess carrier concentration Δn_0 equal to

$$\Delta n_0 = g\tau_n = Jk\beta\tau_n. \quad (6.48)$$

Since the excess carriers have practically the same mobility as the equilibrium carriers, the stationary (steady-state) photoconductivity will be

$$\sigma_{\text{ph0}} = q\beta kJ u_n \tau_n. \quad (6.49)$$

It follows from Eq. (6.49) that the stationary photoconductivity of a semiconductor and, consequently, the photosensitivity of semiconductor radiation detectors is proportional to the excess carrier lifetime τ_n . From this point of view, it is advantageous to use materials with the highest possible τ_n . However, this may substantially increase the time lag of the photodetector.

Indeed, consider the pattern of photoconductivity decay after the light source had been turned off (Figure 6.19). The recombination process reduces the number of excess carriers in compliance with the law [see Eq. (5.50)]

$$\Delta n = \Delta n_0 e^{-t/\tau_n}.$$

The same will be true for the semiconductor's photoconductivity decay (curve

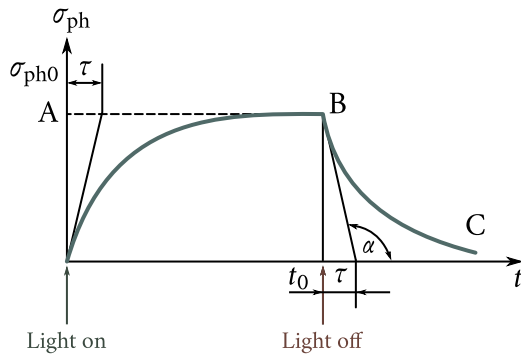


Figure 6.19: Rise in photoconductivity of a semiconductor illuminated by light and photoconductivity decay after illumination has ceased.

BC):

$$\sigma_{ph} = \sigma_{ph0} e^{-t/\tau_n}. \quad (6.50)$$

It follows from Eq. (6.50) that the greater the excess carrier lifetime τ_n the slower the photoconductivity decay rate and, consequently, the greater will be the radiation detector's time lag.

It may easily be demonstrated that the tangent drawn to the photoconductivity decay curve $\sigma_{ph}(t)$ at point t_0 cuts off a section numerically, equal to τ_n , the excess carrier lifetime. This method is often used for determining τ_n .

Figure 6.19 also shows the pattern of the rise in photoconductivity after the semiconductor had been illuminated by a light pulse (curve 0B). The photoconductivity rises gradually and reaches the "plateau" only after a lapse of some time. In this case too, a tangent to the curve $\sigma_{ph}(t)$ drawn at the origin cuts off a section of the straight line AB equal to τ_n .

Excitons. In the act of photoconductivity the electrons from the valence band are transported to the conduction band and become free electrons. However, the process may take another course when the excited electron does not tear its connections with its counterpart hole in the valence band but forms an integral system with it. Ya. I. Frenkel proposed the term *exciton* for such a system. The exciton is similar to an excited hydrogen atom—in both cases there is an electron moving about a unit positive charge and the energy spectrum is a discrete one (Figure 6.20). The exciton levels are near the bottom of the conduction band. Since the excitons are electrically neutral, their appearance does not result in the generation of additional charge carriers. Because of that, the absorption of light is not accompanied by photoconductivity. The present point of view is that, the formation of excitons can in some cases result in photoconductivity. The generated excitons for some

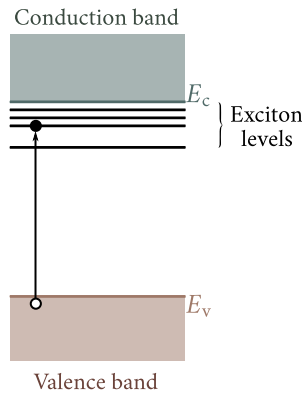


Figure 6.20: Exciton states in a semiconductor.

time wander through the crystal. Colliding with phonons, impurity centres, or other lattice imperfections, the excitons may either recombine or “decompose”. In the first case, the ground state is restored, the excitation energy being transmitted to the lattice or emitted in the form of light quanta (luminescence). In the second case, a pair of free carriers, an electron and a hole, is created. They are responsible for the photoconductivity.

Temperature greatly affects the photoconductivity of semiconductors. As the temperature decreases the number of dark carriers drops. The result is, firstly, an increase in the ratio of photoconductivity to the total conductivity and, secondly, an absolute increase in photoconductivity due to the decrease in the photocarrier recombination rate brought about by the decrease in the dark carrier concentration (the latter effect is observed only in semiconductors with prevailing direct recombination).

Photoresistors. The photoconductivity effect in some semiconductors is widely utilized in photoresistors. Figure 6.21 shows schematically one of the types of photoresistors. It consists of a thin semiconducting film 2 deposited on an insulating substrate 1, of metal electrodes 3, by means of which the photoresistor is connected into a circuit, and of a protective organic film 4. The most sensitive photoresistors are made of cadmium sulfide (CdS), the photoconductivity of which is 10^5 - 10^6 times higher than the dark conductivity. Also, in wide use are photoresistors made of lead sulfide (PbS), which are sensitive to the far infrared radiation. Other semiconducting materials are also being used.

The main advantage of the photoresistors over the vacuum photocells is their high light sensitivity. For instance, the sensitivity of selenium-cadmium photoresistors is 10^5 times higher than that of vacuum photocells. A disadvantage of the

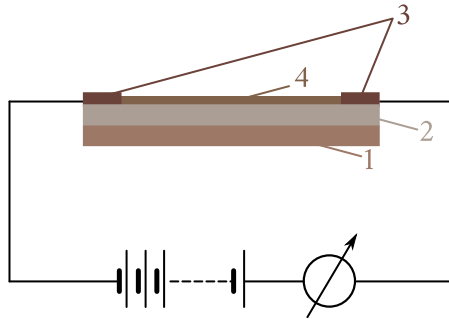


Figure 6.21: Schematic representation of a photoresistor: 1—insulating substrate, 2—semiconducting film, 3—metal electrodes, 4—protective coating.

photoresistors is their time lag.

Electrophotography. The internal photoeffect in semiconductors is widely used in so-called *electrophotography* or *xerography*. The essence of this process is as follows. A thin film of high resistivity semiconductor (usually ZnO) is deposited on a sheet of paper. Before the photographic process the film is negatively charged by a gas discharge. When an image to be photographed is projected onto such paper, the surface charge from the illuminated parts leaks through the film much more readily than from the nonilluminated parts and accordingly an electric image of the object remains on paper after the exposition. To develop the electrical image the paper is sprayed by a weak spray of special dry paint, or “toner”. The particles of toner are deposited on the negatively charged parts of the paper thus developing the image.

The image is fixed by heating the paper to the temperature at which the toner particles melt and adhere firmly to the paper. The main advantage of electrophotography over normal photography is the exclusion of chemical development and fixation processes. This makes it possible to increase the speed of the photographic process drastically, reducing the necessary time down to about ten seconds. However, as yet electrophotography is inferior to normal photography in accuracy and fineness of reproduction and, because of that, its application is limited to cases when great accuracy is not needed (for instance, for multiplying printed texts, cards, etc.). The well known Soviet-made duplicator “Era” operates on this principle.

Semiconductor counters. Apart from light, the internal photoeffect can be excited by irradiating the semiconductor with particles—with electrons, ions, α -particles, etc. Such particles passing through the semiconductor, generate free charge carriers and thus increase its conductivity or the current in a closed constant-voltage circuit. Since the number of generated carriers is proportional to the num-

ber of particles which enter the semiconductor, that number can be determined from the changes in the current. This fact enables semiconductor counters to be devised. Such counters are usually graduated not in units of current but directly in the number of particles. To enhance the sensitivity of the counter the variations of current flowing through the semiconductor are amplified with the aid of special electronic devices.

Semiconductor counters are now in a state of high perfection. They are widely used in nuclear research, in space technology, in medicine, in dosimetry, etc. They will probably play the leading role in radiation detection and spectrometry.

§ 61. Luminescence

A heated body radiates energy the power and the spectral composition of which depend on the temperature of the body. This radiation is termed *thermal*. Its main feature is that it is an equilibrium process. If we place a heated body into a cavity with walls of ideal reflectivity, a dynamical equilibrium is established between the atoms radiating energy and the radiation filling the cavity such that the number of atoms radiating energy and returning to the nonexcited state per unit time would be equal to the number of atoms absorbing radiation and going over to the excited state. This equilibrium can be maintained any length of time. Practically the same equilibrium radiation will be radiated by a heated body which is not surrounded by reflecting walls of a cavity if its temperature is held constant at the expense of energy supplied to it.

Bodies can be made to emit light not only by means of heating. Some materials emit light after they have been irradiated with visible or ultraviolet light, with X-rays, γ -rays, electrons, or other particles, when placed in an electric field, etc. The emitted light may be in the visible part of the spectrum, although the temperature of the emitting body is low (room temperature and below). Such cold emission of light is termed *luminescence* and the bodies exhibiting it are termed *luminophors*; the luminescence excited by light is termed *photoluminescence*. In contrast to thermal radiation, luminescent radiation is a nonequilibrium process. Should a luminescent body be placed in a cavity with reflecting walls, it would loose energy by radiation because the radiated energy reflected by the walls would be absorbed by the body and entirely transformed into the energy of thermal vibrations of its atoms. Therefore, luminescence would eventually cease and the entire energy accumulated in the excited luminophor would be transformed into heat.

The second important feature of luminescence is its long duration in comparison with the period of optic oscillations equal to 10^{-13} - 10^{-15} s. The emission of light in the process of luminescence continues at least 10^{-10} s after the excitation

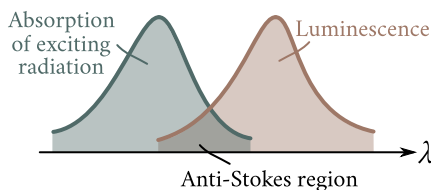


Figure 6.22: Illustration of Stoke's law.

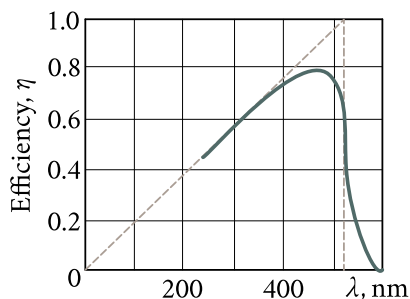


Figure 6.23: Illustration of Vavilov's law.

has ceased. In some instances the emission of light may continue for seconds, minutes, hours and even months after the excitation has ceased. In accordance with the duration of light emission, photoluminescence is divided into phosphorescence and fluorescence. Luminescence with a duration of under 10^{-6} s is usually termed *fluorescence* and that with duration of over 10^{-5} s to 10^{-6} s is termed *phosphorescence*.

The first quantitative investigation of luminescence was undertaken some 100 years ago by Sir George Gabriel Stokes. He succeeded in formulating the following rule which bears his name: the wavelength of light emitted in luminescence is longer than the wavelength of the absorbed light (Figure 6.22). Subsequent experiments have proved that *anti-Stokes* luminescence is also possible when the wavelength of luminescence is shorter than that of the excitation.

An important characteristic of luminescence is its efficiency η , first introduced by S. I. Vavilov. Efficiency is the ratio of the total energy emitted by a body in the process of luminescence to the energy absorbed by the body in the process of excitation. Figure 6.23 shows the dependence of η on the wavelength of excitation. Inside some wavelength interval, the efficiency of luminescence rises in proportion to the wavelength but then drops drastically to zero. This rule was established by S. I. Vavilov and is known as Vavilov's law. The absolute value of the efficiency may be as high as 80% or more.

Let us now discuss the mechanism of luminescence of solid crystals. Experiment shows that crystals with perfect lattice are practically incapable of luminescence. To make them exhibit luminescent properties defects should be created in their structure. The most effective defects are impurity atoms. Such impurities are termed *activators*. Their contents in the matrix material hardly exceeds 10^{-4} . Materials widely used at present are the so-called *phosphor crystals*—complex synthetic crystals with a defect structure possessing high luminescent properties. A phosphor crystal usually contains three components: the matrix, the activator and, the

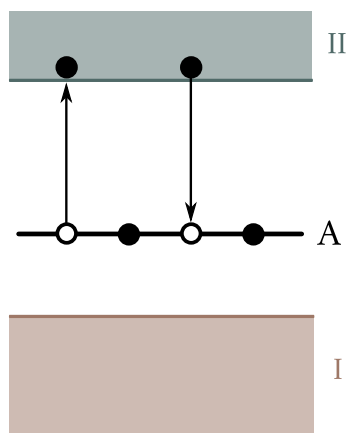


Figure 6.24: Energy diagram of fluorescent luminophor.

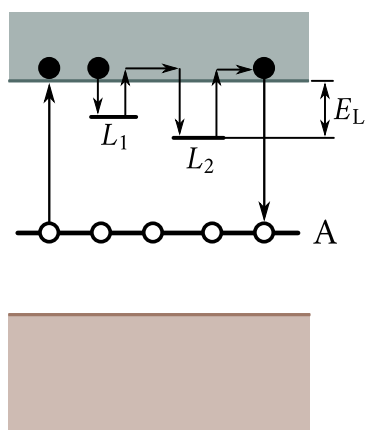


Figure 6.25: Diagram of electronic transitions in the act of phosphorescence.

solvent. The materials frequently used as matrix materials are ZnS, CdS, CaS, etc.; as activators, the heavy metals Ag, Cu, Bi, Mn, etc., and as solvents, the low-melting salts. The spectral composition and the efficiency of luminescence depend both on the matrix material and the activator.

Figure 6.24 shows the energy band pattern of a fluorescent luminophor. There are impurity levels of the activator, A, between the filled band I and the vacant band II. When the activator atom absorbs a photon $h\nu$, an electron from the impurity level A is transported to the conduction band II. As a free electron it wanders freely in the volume of the crystal until it meets an activator ion and recombines with it returning to the impurity level A. The recombination is accompanied by the emission of a quantum of fluorescent light. The decay time of luminescence of a luminophor is determined by the lifetime of the excited state of the activator atoms, which seldom exceeds 10^{-9} s. Therefore, fluorescence is a short-lived process and terminates almost immediately after the irradiation of the body has ceased.

For durable luminescence characteristic of phosphorescence the luminophor must contain not only the activator atoms A but also electron traps L near the bottom of the conduction band (Figure 6.25). Such traps may be formed by impurity atoms, interstitial atoms, vacancies, etc. The light absorbed by the luminophor excited the activator atoms: the electrons from the impurity level A go over to the band II and become free electrons. Trapped by traps they lose their mobility together with the ability to recombine with the activator ions. To liberate an electron from the trap the energy E_L should be expended. This energy the electron can obtain from the lattice vibrations. The time τ spent by the electron in a trapped state

is proportional to $e^{E_L/(k_B T)}$; it may be quite large if E_L is great enough.

The electron that left the trap wanders through the crystal until it is again trapped or recombines with an activator ion. In the latter case, a quantum of luminescent light is emitted. Hence, the traps serve as centres where the energy of absorbed photons is accumulated so as to be subsequently emitted in the form of luminescent light. The duration of this emission is determined by the time the electrons spend in the traps.

Experiments show that not in all cases is the transition of the electron from an excited state to the ground state accompanied by the emission of a light quantum. A much more frequent result is the generation of a phonon. For this reason, the purity of the phosphor crystals must satisfy the most severe requirements. Often a negligible impurity concentration (less than 10^{-4} percent) completely extinguishes the luminescence.

The quantum theory presents a simple explanation of the fundamental laws of luminescence including Stokes' and Vavilov's laws.

Stokes' law. When a luminophor is irradiated by light quanta, the energy of the quanta is partly spent on the excitation of the activator atoms and partly is transformed into energy of other types (mainly heat). Denote the fraction of the quantum's energy spent on exciting the activator atom by ε . When the atom returns from an excited to the ground state, a quantum of luminescent light will be emitted with its energy equal, evidently, to ε . The corresponding frequency is $\nu = \varepsilon/h$ and the wavelength is $\lambda = hc/\varepsilon$. Since the energy of the incident quantum $\varepsilon_0 > \varepsilon$, the wavelength λ of the luminescent light should be longer than that of the light which initiates luminescence ($\lambda > \lambda_0$) and this is what Stokes' law states.

When the incident quantum collides with an excited atom, its energy $\varepsilon_0 = h\nu_0$ may be added to the excitation energy ε causing the generation of a quantum with energy exceeding the energy of the one which initiates luminescence. This is the origin of *anti-Stokes luminescence*.

Vavilov's law. Consider the simplest case of every incident photon $\varepsilon_0 = h\nu_0$ generating a luminescent photon $\varepsilon = h\nu$ (quantum efficiency unity). Then, the efficiency of the luminescence will evidently be equal to the ratio of the energies of those photons: $\eta = \varepsilon/\varepsilon_0$. Since $\varepsilon = h\nu = hc/\lambda$, it follows that

$$\eta = \frac{\nu}{\nu_0} = \frac{\lambda_0}{\lambda}. \quad (6.51)$$

From Eq. (6.51), we see that the luminescence efficiency should grow in proportion to the wavelength of the excitation, as required by Vavilov's law. When λ_0 attains a value for which the energy of the incident quanta is not enough to initiate luminescence, the efficiency drops abruptly to zero.

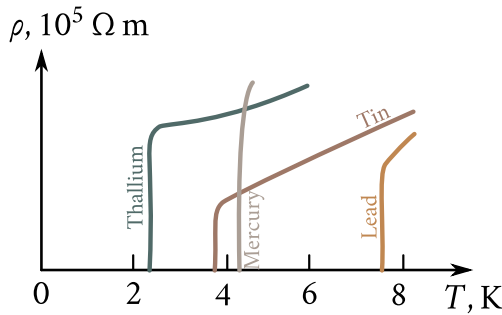


Figure 6.26: Abrupt change in specific resistance of mercury, tin, lead and thallium in the course of transition to superconducting state.

§ 62. Fundamentals of superconductivity

Phenomenon of superconductivity. Investigating the role played by impurities in residual resistance, H. Kamerlingh Onnes in 1911 carried out experiments with ultrapure mercury. The results of those experiments were startling: at a temperature $T_{\text{cr}} = 4.2$ K the specific resistance ρ of mercury fell to zero (Figure 6.26). This phenomenon became known as *superconductivity*. The temperature T_{cr} at which the transition to the superconducting state takes place is termed *critical*, or *transition*, temperature. For thallium, tin and lead it is equal to 2.53 K, 3.73 K and 7.19 K, respectively (Figure 6.26).

Since according to Ohm's law $\rho = \mathcal{E}/i$, the condition $\rho = 0$ means that, for a finite current density i , the intensity of the electric field \mathcal{E} at any point of the conductor is zero: $\mathcal{E} = 0$.

Experiments carried out at M.I.T. showed that a current of several hundred amperes once induced in a superconducting ring continued to flow without attenuation for a whole year.

Up to now over 20 pure chemical elements and several hundred alloys and chemical compounds have been found to be superconductive. They have transition temperatures ranging from 0.01 K to 20 K.

In 1933, W. Meissner and R. Ochsenfeld, found that the phenomenon of superconductivity consists not only in ideal conductivity, that is, zero specific resistance. The magnetic field is pushed out of the bulk of a superconductor no matter how this field was established—by an external magnet or by a current flowing in the superconductor itself. This means that the magnetic induction B_i inside the superconductor is always zero as long as it is in the superconducting state.

In other words, the superconductor is an ideal diamagnetic whose magnetic susceptibility $\chi = -1$. It will be shown in the following chapter that normal dia-

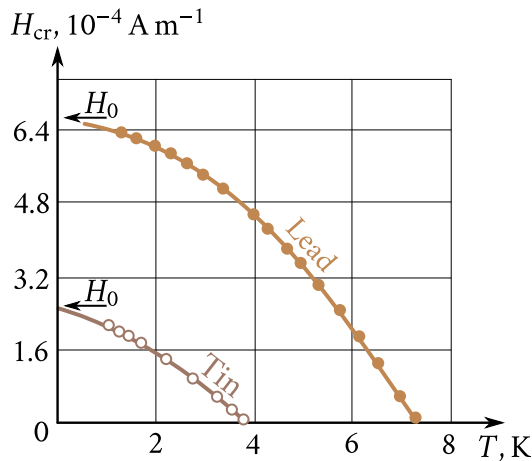


Figure 6.27: Temperature dependence of critical field intensity H_{cr} in superconductor.

magnetics have $|\chi| \ll 1$.

Hence, superconductivity is a combination of two simultaneous phenomena—that of ideal conductivity and of ideal diamagnetism.

The superconductive state can be destroyed by a magnetic field. The necessary magnetic field H_{cr} is termed *critical*. The value of H_{cr} depends on the temperature: at $T = T_{cr}$, the critical field intensity is zero. With the decrease in temperature H_{cr} rises and is maximal at absolute zero. The temperature dependence of H_{cr} for lead and tin is shown in Figure 6.27.

Fundamentals of theory of superconductivity. Despite the fact that over 60 years have gone by since superconductivity was first discovered, the microscopic theory of this phenomenon is a quite recent development due mainly to Bardeen, Cooper, Schrieffer. The abbreviation for it is the BCS theory. Let us discuss this theory in general terms.

Gap in the energy spectrum of conduction electrons in a superconductor. We again recall the causes of a finite electrical resistance of normal conductors, for instance, of metals in a normal state. If we neglect the periodic nature of the metal's lattice potential (the *Sommerfeld model*), we can regard it as a potential trough for the electrons, having a flat bottom and filled by electrons up to the Fermi level E_F (see Figure 3.4). The kinetic energy of such electrons is given by Eq. (5.11)

$$E = \frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m}.$$

Figure 6.28(a) shows once again the E versus k dependence corresponding to Eq. (5.11): thin horizontal lines denote the occupied levels, the solid line denotes the

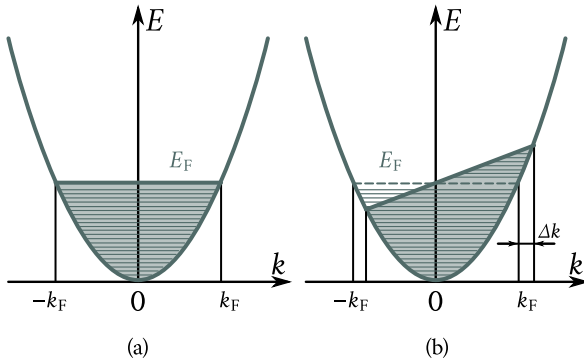


Figure 6.28: Dependence of free electron energy in conduction band of a metal on wave vector: (a)—in the absence of external field; (b)—external field \mathcal{E} imparts additional momentum to electrons, increasing their wave vector by Δk .

Fermi level E_F , and k_F and $-k_F$ are wave vectors corresponding to this level.

The application of an electric field \mathcal{E} causes a change in the electron distribution over the states [see Figure 6.1(a)]; the electrons are transported from the left-hand side to the right-hand side. In Figure 6.28(b), this corresponds to the transition of the electrons from the region of negative k 's to the region of positive k 's. Such transitions are possible since there is a practically unlimited number of unoccupied states above the Fermi level which the electrons can occupy.

Should there be no limiting factors, the momentum of the conduction electron would in time Δt grow¹ under the influence of the field \mathcal{E} by an amount $p_{\mathcal{E}} = \hbar \Delta k = q \mathcal{E} \Delta t$, and a current of a density $i = q n v p_{\mathcal{E}} = q n / (q \mathcal{E} / m) \Delta t = (q^2 n / m) \mathcal{E} \Delta t$ would be established in the conductor, its magnitude growing infinitely with time. This would correspond to infinite specific conductance of the conductor, since

$$\lim_{\Delta t \rightarrow \infty} \sigma = \lim_{\Delta t \rightarrow \infty} \frac{q^2 n}{m} \Delta t \rightarrow \infty. \quad (6.52)$$

However, should it even be possible to realize condition (6.52), this would still not amount to ideal conductivity, which is characterized, as we have seen, by the condition that the current density for $\mathcal{E} = 0$ is finite: $i \neq 0$. But the condition (6.52) cannot be realized in any case. The factors that prevent this are the processes of electron scattering by lattice defects and, primarily, by thermal lattice vibrations—phonons—which are present down to absolute zero. Here, the main part is played by elastic scattering processes which change the electron's momentum to one directly opposite, so that they move over from the right-hand side of the distribu-

¹The product $q \mathcal{E} = F$ is the force with which field \mathcal{E} acts on the electron. If Δt is a time interval, $F \Delta t$ is the impulse of force.

tion curve to the left-hand side. The corresponding transitions in Figure 6.28(b) are those from the region of positive k 's to the region of negative k 's. The rate of the scattering processes is the greater the greater the field disturbs the equilibrium distribution of the electrons over the states, that is, the greater the displacement to the right of the distribution curve shown in Figure 6.1(a) by a dotted line. Those processes bring the electron drift velocity down to the value $v_d = q\mathcal{E}\tau_F/m$, the current density to $i = q^2 n \mathcal{E} \tau_F / m$, and the specific conductance to $\sigma = q^2 n \tau_F / m$, where τ_F is the relaxation time of the electrons occupying levels close to the Fermi level.

Let us make the following point important for the future. At least two conditions should be fulfilled to make elastic transitions, which are responsible for finite electrical resistance of a normal metal, possible: (a) there should be states the scattered electrons can occupy (in other words, the corresponding energy levels should lie in the allowed energy band); (b) the states the scattered electrons are to occupy must not already be occupied.

For a normal metal with an energy spectrum of conduction electrons as shown in Figure 6.28, both those conditions are fulfilled making the scattering processes possible.

Can a model of the energy spectrum of conduction electrons be built which would make scattering processes (at least under certain conditions) impossible even in the presence of scattering centres—phonons, impurity atoms, etc.?

Apparently, yes. Such a spectrum is shown in Figure 6.29(a). The difference between it and the spectrum shown in Figure 6.28 is that in it there is an energy gap $E_{e.g}$ with the Fermi level E_F in the middle. The lower part of the conduction band is completely occupied by electrons; the upper part above the gap is completely free. The band pattern looks like that of an intrinsic semiconductor at $T = 0$ K whose specific conductance in case of such occupation of the bands is zero. Since the metal retains its high specific conductance at $T \approx 0$ K, it should be presumed that in contrast to a semiconductor whose energy gap (the forbidden band) does not change its position in an external field, the $E_{e.g}$ in the conduction band of the metal moves in the electric field together with the electron distribution, as shown in Figure 6.29(b). During a finite time interval Δt , the field \mathcal{E} exists in the superconductor, the electron's wave vector increases by an amount $\Delta k = p_{\mathcal{E}}/\hbar = q\mathcal{E}\Delta t/\hbar$ and the energy gap $E_{e.g}$ shifts to the right together with the electron distribution by a distance Δk .

Now let us consider the possibility for the electron q occupying the upper level of the right-hand subband to be scattered. The arrows 1, 2, 3 in Figure 6.29(b) show the possible scattering processes: 1 is elastic scattering resulting in the change from k to $-k$; 2 are transitions to the levels of the lower left-hand subband; 3 are transi-

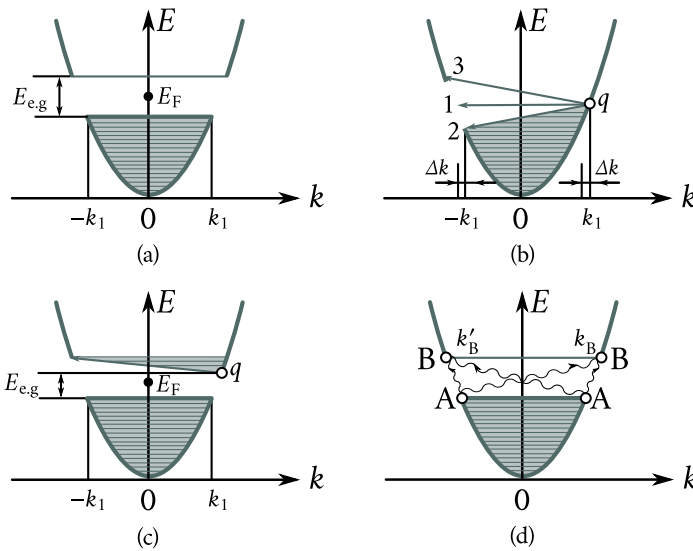


Figure 6.29: Energy spectrum of conduction electrons in metal with a mobile energy gap (explanation in text).

tions to the levels of the upper left-hand subband. It may easily be seen that transitions 1 are forbidden since they terminate in the forbidden part of the energy spectrum, namely $E_{e.g.}$. Transitions 2 are forbidden by the Pauli exclusion principle, since the corresponding levels are already occupied by electrons. Transitions 3, although allowed, require an ionization energy equal to $E_{e.g.}$. If the metal's temperature is low enough so that the mean phonon energy $\hbar\omega_{ph} < E_{e.g.}$, those transitions are impossible².

²To be exact, there are phonons with an energy $\hbar\omega_{ph} > E_{e.g.}$ capable of exciting electrons from the lower subband to the states of the upper subband even at the lowest temperature. This should cause the appearance of vacant levels in the lower subband and of "normal" electrons in the upper subband. One would think that the appearance of vacant levels in the lower subband would bring about the scattering of electrons responsible for ideal conductivity; in other words, that it would in effect destroy superconductivity. Actually, as a more detailed consideration shows, a rise in temperature is accompanied by a narrowing of the energy gap [Figure 6.29(c)] so that no vacant level capable of scattering the electrons of the lower subband remains in it. In other words, phonons with an energy $\hbar\omega_{ph} > E_{e.g.}$ not only transform a superconducting electron A into a normal one B' but destroy the superconducting state of this electron and of the electron A', which was paired with A, by exciting it to the normal state [Figure 6.29(d)]. As temperature rises the number of "energetic" phonons decreases, the width of the energy gap decreases together with the number of superconducting electrons. On the contrary, the number of normal electrons rises. At $T = T_{cr}$, the width of the gap vanishes (Figure 6.34), all electrons go over to the normal state, and superconductivity is destroyed.

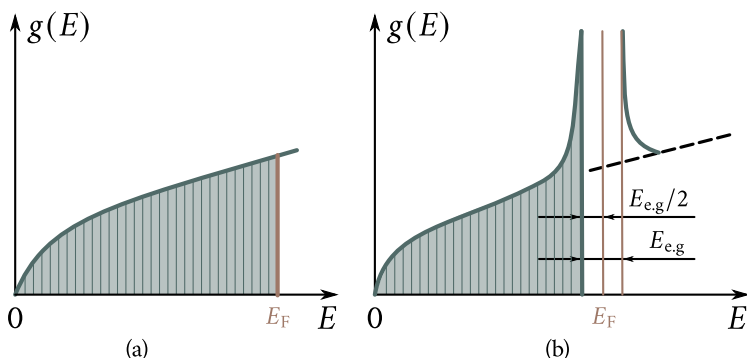


Figure 6.30: Variation of density of states of free electrons in metal with the appearance of an energy gap in the conduction band: (a)—density of states versus energy plot for a normal metal; (b)—ditto for a metal with a gap in its conduction band.

Hence, there are conditions in which even in the presence of such scattering centres as phonons, the scattering processes limiting the conductivity in a metal, whose electron energy spectrum has a “mobile gap” (Figure 6.29), cannot take place. Accordingly, such a metal may become an ideal conductor, just like a superconductor.

Let us look again at Figure 6.29(a). The tangent to the curve $E(k)$ near the top of the lower, filled, part of the conduction band runs horizontally ($dE/dk = 0$) and, accordingly, the translational velocity of the electrons occupying these levels $v = \hbar^{-1}(dE/dk) = 0$, although their momentum p_1 and wave vector $k_1 = p_1/\hbar$ are quite large. We encountered a similar model when we discussed semiconductors (see Figure 5.12); this property of the electrons will prove to be essential in constructing the model of superconductivity.

Figure 6.30(a) shows the dependence of the density of states in the conduction band of a normal metal on energy for $T = 0$ K and Figure 6.30(b) the pattern of this dependence in the presence of a gap $E_{e.g.}$ in the conduction band. Near the edges of the gap the density of the states is higher and because of that the band made shorter by $E_{e.g.}/2$ still has enough states to accept all the electrons of the conduction band.

Hence, if we were to prove that metals can actually have an electron energy spectrum with a “gap” and if the causes of its appearance could be established, the miracle of the ideal conductivity of superconductors would generally be unveiled. For this reason, the efforts of investigators in superconductivity were concentrated on the experimental verification of the presence of such a gap in the energy spectrum of superconducting metals.

At present, a number of methods have been devised capable not only of detecting the gap but also of measuring its width. One of them is based on the study of

far infrared radiation absorption by metals. The idea of the method is as follows. Should a superconductor be irradiated with electromagnetic radiation of continuously varying frequency ω , it would not be absorbed as long as the energy of its quantum remained less than the width of the gap $E_{\text{e.g}}$ (of course, if there is such a gap). Intense absorption should start at a frequency ω_{cr} , for which $\hbar\omega_{\text{cr}} = E_{\text{e.g}}$, increasing with the frequency to values common to a normal metal. Measuring ω_{cr} one can determine $E_{\text{e.g}}$.

The experiments convincingly proved that there is a gap in the electron energy spectrum of superconductors. Table 6.6 shows the values of the gap width at $T = 0$ K for some metals together with the transition temperature. We see that the gap $E_{\text{e.g}}$ is very narrow, approximately 10^{-3} eV to 10^{-2} eV wide, and that there is a direct connection between the gap's width and the transition temperature T_{cr} ; the higher the T_{cr} the greater the $E_{\text{e.g}}$ is.

After the presence of a gap in the energy spectrum of conduction electrons in superconductors was proved experimentally, attention turned to the problem of the origin of this gap.

Electron pair formation. As we already know, the formation of forbidden bands in the energy spectrum of semiconductors is due to the interaction of the electrons with the periodic field of the crystal lattice.

It would be natural to suppose that the energy gap in the conduction band of a metal in the superconducting state is also due to some additional electronic interaction that appears when the metal enters that state. Its origin is as follows.

A free electron moving through the lattice interacts with the ions, “pulling” them from their equilibrium sites (Figure 6.31) and creating an excess positive charge that may attract another electron. For this reason, apart from the usual Coulomb repulsion, a force of attraction can arise between the electrons owing to the presence of the positive ion lattice. If this force of attraction exceeds the force of repulsion, it will be advantageous from the viewpoint of energy for the electrons to associate into couples. These became known as *Cooper pairs*.

Table 6.6

	Al	Sn	Hg	V	Pb	Nb
$E_{\text{e.g}}(0)$ (10^3 eV)	3.26	11.0	16.4	14.3	21.4	22.4
T_{cr} (K)	1.2	3.73	4.15	4.9	7.19	9.22
$E_{\text{e.g}} = 3.5k_{\text{B}}T_{\text{cr}}$ (10^3 eV)	3.6	11.2	12.5	14.8	21.7	27.7

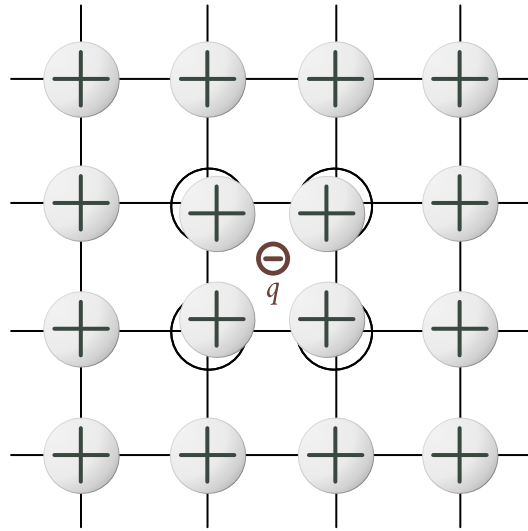


Figure 6.31: Moving electron polarizes the lattice and pulls positive ions a little away from their equilibrium sites, thereby creating an excess positive charge which attracts another electron so that it forms an electron pair with the first electron.

The formation of a Cooper pair results in the reduction in the energy of the two electrons by the amount equal to the binding energy of the electrons in the pair, E_b . This means that a conduction electron, which in a normal metal had a maximum energy E_F at $T = 0$ K [see Figure 6.28(a)], in the superconducting state has an energy $E_b/2$ less (the energy per pair being E_b less) since this is the energy that must be spent to break up the pair and move the electrons to the normal state. Therefore, in the one-electron spectrum there must be a gap of $E_{e.g} = 2E_b$ between the upper level of a coupled electron and the lower level corresponding to the normal state, which is required for superconductivity. It may easily be seen that this gap is mobile, that is, it can shift in an external field together with the electron distribution curve over the states.

Figure 6.32 is a schematic representation of a Cooper pair. It consists of two electrons oscillating about the induced positive charge, which in some ways resembles a helium atom. Each electron of the pair may have a large momentum p_F and a large wave vector k_F ; the pair as a whole (its centre of masses), on the other hand, can remain stationary having zero translational velocity. This explains a paradoxical property of the electrons occupying the upper levels of the filled part of the conduction band in the presence of a gap [see Figure 6.29(a)]. Such electrons have very large p 's and k 's ($p \approx p_F$ and $k \approx k_F$) and a translational velocity $v \propto dE/dk = 0$. Since the central positive charge is induced by the moving elec-

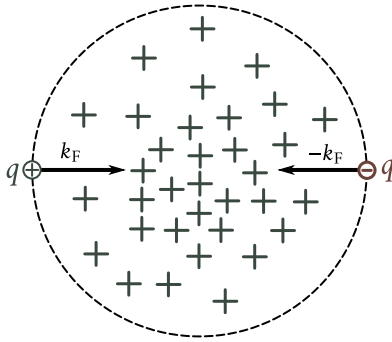


Figure 6.32: Schematic model of a Cooper pair.

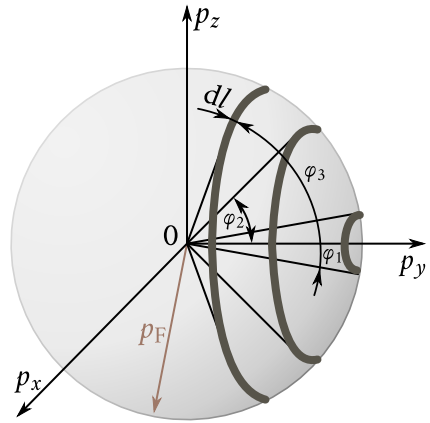


Figure 6.33: Estimating the number of electrons capable of forming Cooper pairs.

trons themselves, the Cooper pair acted upon by an external field can freely drift in the crystal, the energy gap moving with the electron distribution, as shown in Figure 6.29(b). Hence, the conditions for superconductivity are fulfilled from this point of view as well.

However, not all the conduction electrons can form Cooper pairs. Since this process involves a change in energy, only electrons that can change their energy can form pairs. Only the electrons in a narrow strip close to the Fermi level (the Fermi electrons) are free from this limitation. A rough estimate gives the fraction of such electrons as being approximately 10^{-4} of the total number of electrons and the width of the strip of the order of $10^{-4}p_F$.

Figure 6.33 shows a Fermi sphere of a radius p_F in momentum space. Rings dl wide make angles $\varphi_1, \varphi_2, \varphi_3$ with the p_y axis. The electrons the ends of whose momentum vectors \mathbf{p}_F lie within the area of a given ring, make up a group every member of which has the same absolute value of momentum p_F . The number of electrons in each group is proportional to the area of the respective ring. Since the area of the ring rises with φ so does the number of electrons in the band. Electrons of any group may form pairs, but the maximum number of pairs will be formed by the electrons of the more numerous group. The latter is made up of electrons whose momenta are equal in magnitude and opposite in direction. The ends of the vectors \mathbf{p}_F of such electrons are not limited to a narrow band but spread over the entire Fermi surface. Those electrons are so numerous in comparison with any other electrons that practically only one group of Cooper pairs is formed—that made up of electrons whose momenta are equal in magnitude and opposite in direction.

A remarkable peculiarity of such pairs is the ordering of their momenta: the centres of masses of all the pairs have identical momenta, being zero when the pairs are at rest and nonzero when the pairs move in the crystal. The result is a rather rigid correlation between the motion of every single electron and the motion of all the other electrons bound into pairs.

The electrons “move like mountain-climbers tied together by a rope: should one of them leave the ranks due to the irregularities of the terrain (caused by the thermal vibrations of the lattice atoms) his neighbours would pull him back”³. This property makes the ensemble of Cooper pairs little susceptible to scattering. Accordingly, should the pairs acted upon by some external force be set in motion, the current established by them would continue to flow in the superconductor indefinitely even after the factor that brought it to life ceased to operate. Since only the electric field \mathcal{E} can play the role of such a factor, this means that in a metal, in which Fermi electrons are bonded into Cooper pairs, a once excited electric current i can remain unaltered even after the field has vanished: $i = \text{constant}$ at $\mathcal{E} = 0$. This proves that the metal is actually in the superconducting state and that its conductivity is ideal. Such state of the electrons may be roughly compared with the state of a body moving without friction: the body having received a momentum can move indefinitely and its momentum remains constant.

In the above, we compared a Cooper pair to a helium atom. However, such a comparison should be treated very cautiously. As was already stated, the positive charge is not exactly constant and stationary, as in the case of a helium atom, but is induced by the moving electrons themselves and moves with them. Moreover, the binding energy of the electrons in a pair is many orders of magnitude less than their binding energy in the helium atom. According to Table 6.6 the binding energy of Cooper pairs $E_b = 10^{-3}$ eV to 10^{-2} eV, the corresponding value for the helium atom being $E_b = 24.6$ eV. Because of that, the dimensions of the Cooper pair are many orders of magnitude larger than that of the helium atom. Calculations show the effective diameter of a pair to be $L \approx 10^{-7}$ m to 10^{-6} m; another term for it is *coherence length*. There are about 10^6 centres of masses of Cooper pairs inside the effective volume L^3 of one such pair. For this reason, such pairs cannot be regarded as separately existing “quasimolecules”. On the other hand, the accompanying colossal overlapping of the wave functions of numerous pairs enhances the electron pairing effect so that it manifests itself in macroscopic proportions⁴.

³Ya. I. Frenkel: *Introduction to the Theory of Metals*, GITTL, Moscow (1950) (in Russian).

⁴There is another analogy, a very profound one at that, between a Cooper pair and a helium atom. The essence of it is that an electron pair constitutes a system with integral spin, the same as the helium atom ${}^4\text{He}$ does. It is a known fact that the superfluidity of helium may be considered as the result of a peculiar effect of bosons condensing on the lowest energy level. From this point of

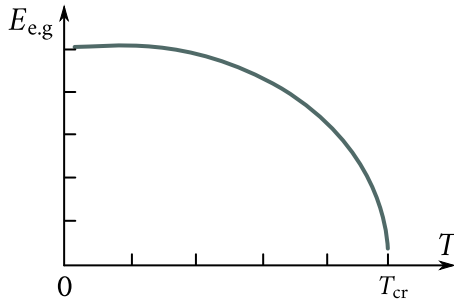


Figure 6.34: Variation of the energy gap $E_{e,g}(T)$.

Hence, electron pairing is a typical collective effect. There are no forces of attraction acting between two isolated electrons to make their coupling possible. In effect, the entire ensemble of the Fermi electrons together with the lattice atoms takes part in the formation of a pair. Because of that, the binding energy (the gap width $E_{e,g}$) too depends on the state of the electron-atom ensemble as a whole. At absolute zero, when all the Fermi electrons are in pairs, the width of the gap is at its maximum, $E_{e,g}(0)$. The rise in temperature is accompanied by the generation of phonons capable in the act of scattering of transmitting energy to the electrons sufficient to break up the pair. At low temperatures the concentration of such phonons is not large and the breaking up of a pair is a rare event. The disappearance of some pairs cannot, naturally, lead to the disappearance of the gap for the remaining pairs but makes it somewhat narrower, with the edges of the gap drawing closer to the Fermi level [see Figure 6.29(c)]. With a further rise in temperature the phonon concentration grows very rapidly, their mean energy growing as well. The result is a steep rise in the break-up rate of the pairs and, accordingly, a drastic decrease in the gap width for the remaining pairs. At some temperature T_{cr} the gap disappears altogether (Figure 6.34), its edges merging with the Fermi level and the metal returning to the normal state. The temperature T_{cr} is just the critical transition temperature that was mentioned at the beginning of the section.

It follows then that the critical temperature for the transition of a metal to the superconductive state should be the greater the greater the gap width at absolute zero, $E_{e,g}(0)$, is. The BCS theory gives the following approximate dependence of

view, superconductivity may be regarded as superfluidity of the Cooper electron pairs. The analogy is a still more far-reaching one. Another helium isotope ^3He , whose nucleus has a half-integral spin, does not exhibit superfluidity. But a most striking new discovery is that in the lowest temperature range the atoms of ^3He can form pairs quite like the Cooper pairs and the liquid becomes superfluid. One is justified in saying that the superfluidity of ^3He is a sort of superconductivity of its atomic pairs.

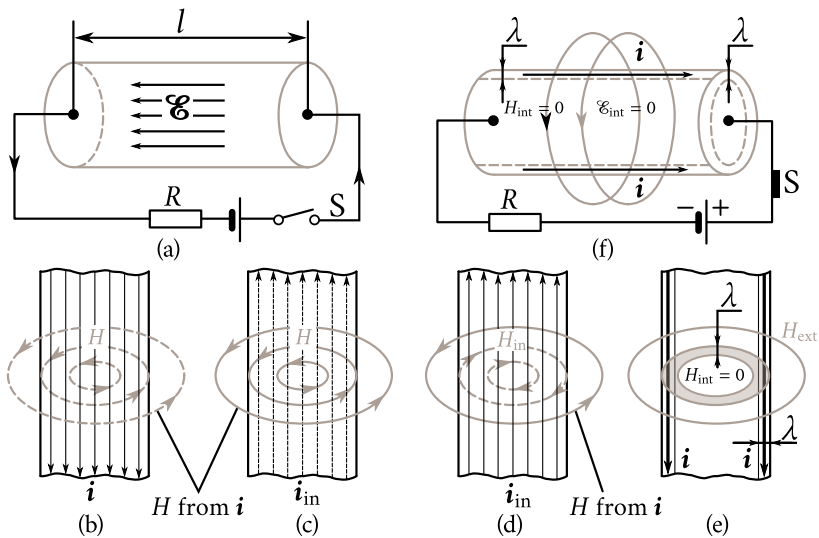


Figure 6.35: Behaviour of superconductor connected into an electric circuit (see text).

$E_{e.g}(0)$ on T_{cr} :

$$E_{e.g}(0) = 3.5k_B T_{cr} \quad (6.53)$$

which is in good agreement with experiment (see the last line in Table 6.6).

Behaviour in an external electric field. Connect a long cylindrical superconducting specimen into an electric circuit, as shown in Figure 6.35(a). As the circuit is closed, a homogeneous electric field is established in the specimen, $\mathcal{E} = V/l$, where V is the voltage across the specimen and l is its length. Acted upon by the field \mathcal{E} all Cooper pairs contained in the specimen will start to move against the field with the same acceleration

$$a = \frac{2q\mathcal{E}(t)}{2m} = \frac{q\mathcal{E}(t)}{m}$$

where $2q$ is the pair's charge, and $2m$ its mass. The current density in the superconductor will start to grow:

$$\frac{di}{dq} = 2q \left(\frac{n_s}{2} \right) \frac{dv_d}{dt} = qn_s a = \frac{q^2 n_s}{m} \mathcal{E}(t)$$

where v_d is the drift velocity of the pairs, $n_s/2$ their number, and n_s is the concentration of "superconducting" electrons.

The current i generates a solenoidal magnetic field H in the superconductor [Figure 6.35(b)]. Since i grows with time so does the magnetic field H . This results in the appearance of an induced electric field \mathcal{E}_{in} in directed against \mathcal{E} and of an induced current i_{in} directed against i [Figure 6.35(c)]. The current i_{in} generates a

magnetic field H_{in} directed against H [Figure 6.35(d)]. The result is the compensation of the field \mathcal{E} inside the superconductor by the field \mathcal{E}_{in} and of the field H by the field H_{in} , so that the resulting electric field in the specimen $\mathcal{E} = 0$ [Figure 6.35(f)] together with the resulting magnetic field $H_{\text{int}} = 0$. For such compensation to continue it is necessary, firstly, that the current of Cooper pairs, i , be maintained in the specimen indefinitely after the end of transient processes. To this end, the specimen's resistance should be zero and this is so if the specimen is in the superconducting state. Secondly, this current should be localized in a thin surface layer λ of the superconductor [Figures 6.35(e,f)], for in this case it does not generate a magnetic field inside the specimen but generates an external field H in the surrounding space just as a normal current does.

Hence, after transient processes come to an end the following stationary state is established in the specimen:

$$\mathcal{E} = 0, \quad i = \text{constant}, \quad H_{\text{int}} = 0.$$

The first two conditions correspond to ideal conductivity and the third to ideal diamagnetism. In the stationary state, Cooper pairs move without acceleration (free motion) with the same drift velocity $v_d = p_{\mathcal{E}}/m$, where $p_{\mathcal{E}}$ is the momentum accumulated by the pair during the time the circuit was closed. The current set up by them is

$$i = 2q \left(\frac{n_s}{2} \right) v_d = q n_s \frac{p_{\mathcal{E}}}{m}.$$

As has already been stated before, this current is localized in a thin surface layer λ of the sample, the magnetic field of the current being concentrated in this layer [Figures 6.35(e,f)]. The parameter λ is termed the *penetration depth*. Theory gives the following expression for this parameter:

$$\lambda = \left(\frac{m}{q^2 n_s \mu_0} \right)^{1/2} \quad (6.54)$$

where μ_0 is the permeability of free space. For different superconductors λ lies in the range from 4×10^{-8} m to 10^{-7} m.

It follows from Eq. (6.54) that at $T = T_{\text{cr}}$, when the concentration of “superconducting” electrons vanishes, λ becomes infinite. Physically, this means that, as the metal returns to the normal state the layer λ in which the magnetic field is localized spreads across the entire cross section of the sample and ideal diamagnetism vanishes.

Behaviour of superconductor in magnetic field. Now let us suppose that a magnetic field H_{ext} is set up in space containing a cylindrical superconducting sample [Figure 6.36(a)]. The field induces a solenoidal electric field in the sample which sets up a solenoidal electric current i_{in} . The current i_{in} creates a magnetic

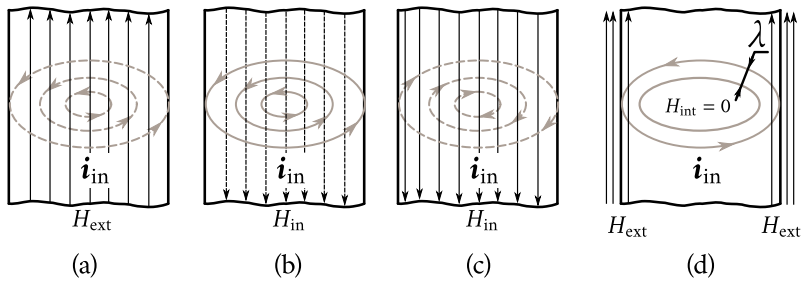


Figure 6.36: Behaviour of superconductor in magnetic field (explanation in text).

field H_{in} [Figure 6.36(b)] directed against the external field and compensates it. The field H_{in} in its turn generates the current i'_{in} which compensates the current i_{in} [Figure 6.36(c)]. The overall effect is the compensation of the external field H_{ext} by the field H_{in} and of the current H_{in} by H'_{in} [Figure 6.36(d)]. The total induced current flows in a thin surface layer λ in which its magnetic field that compensates the external field is localized. After the termination of transient processes the same steady state is established in the sample as was the case when an emf was applied to it:

$$\mathcal{E} = 0, \quad i = \text{constant}, \quad H_{int} = 0.$$

Naturally, this state can be established only if the current i_{in} induced during the time the magnetic field was switched on continues indefinitely, that is, if the sample is in the superconducting state.

Destruction of superconducting state by fields. Before the field is switched on, the momenta of the electrons making up a pair are equal in magnitude and opposite in direction, the momentum of the centre of masses of the pair is zero. Acted upon by the field \mathcal{E} every pair as a whole attains some drift velocity v_d and increases its energy by the amount

$$\Delta E = 2 \left(\frac{mv_d^2}{2} \right).$$

If this energy exceeds the binding energy of the pair $E_b = E_{e.g}/2$, the pairs start to break up and the superconducting state would start to vanish. For this reason, the condition for the transition of a metal from the superconducting to the normal state may be written as follows:

$$2 \left(\frac{mv_d^2}{2} \right) \geq \frac{E_{e.g}}{2}.$$

From here we can easily find the drift velocity

$$v_d = \left(\frac{E_{e.g}}{2m} \right)^{1/2}$$

and the current density

$$i_{cr} = qn_s v_d = qn_s \left(\frac{E_{e.g}}{2m} \right)^{1/2} \quad (6.55)$$

for the case when superconductivity in a metal begins to vanish. Setting $E_{e.g} = 5 \times 10^{-4}$ eV, $q = 1.6 \times 10^{-19}$ C, and $n_s = 10^{18} \text{ cm}^{-3}$, we obtain $v_d = 1.8 \times 10^4 \text{ m s}^{-1}$ and $i_{cr} = 2.5 \times 10^5 \text{ A cm}^{-2}$.

With account taken of the fact that the current i sets up a magnetic field on the surface of the sample of intensity

$$H = i\lambda, \quad (6.56)$$

(λ is the penetration depth of the magnetic field into the superconductor) condition (6.55) may be formulated as follows: the superconducting state of the sample will be destroyed when the intensity of the magnetic field on its surface will attain the following critical value:

$$H_{cr} = i_{cr}\lambda. \quad (6.57)$$

Substituting λ from Eq. (6.54) and i_{cr} from Eq. (6.55) we obtain

$$H_{cr} = \left(n_s \frac{E_{e.g}}{2\mu_0} \right)^{1/2}. \quad (6.58)$$

For $E_{e.g} = 5 \times 10^{-4}$ eV and $n_s = 10^{18} \text{ cm}^{-3}$, $H_{cr} \approx 10^4 \text{ A m}^{-1}$ (or 100 Oe).

Thus, when the magnetic field on a superconductor's surface attains the critical value H_{cr} determined by condition (6.58), the superconductivity is destroyed. Since the gap width substantially depends on temperature [see Eq. (6.53)], i_{cr} and H_{cr} should also depend on temperature, their values decreasing with the rise in temperature. The BCS theory gives the following dependence of H_{cr} and I_{cr} on absolute temperature:

$$H_{cr} = H_{cr}(0) \left[1 - \left(\frac{T}{T_{cr}} \right)^2 \right], \quad (6.59)$$

$$I_{cr} = \pi d H_{cr}, \quad \text{for } d \gg \lambda, \quad (6.60)$$

where $H_{cr}(0)$ is the critical magnetic field intensity at $T = 0$ K, and d is the diameter of the specimen. There is a satisfactory agreement between those relations and experiment.

Practical uses of superconductivity. The field of practical application of superconductivity widens from year to year. First, it serves as a basis for super-

conducting magnets. Such magnets are solenoids or electromagnets with a ferromagnetic core with the winding made of a superconducting material. Calculations show that to establish a magnetic field intensity of $8 \times 10^6 \text{ A m}^{-1}$ ($\approx 10^5$ Oe) in a solenoid of a diameter of one metre, a superconducting magnet requires 10^4 times less power than an ordinary electromagnet. Recently, superconducting magnets using Nb_3Sn have been fabricated which enable magnetic fields up to $6 \times 10^6 \text{ A m}^{-1}$ ($\approx 7 \times 10^4$ Oe) to be produced.

Superconductivity is also being utilized to design modulators (converters of weak constant current into an audio-frequency current), rectifiers for the detection of modulated high frequency oscillations in which the use is made of the nonlinearity of the superconductor's resistance in the transitional region, commutators (noncontact switches utilizing the phenomenon of superconductivity), cryotrons (superconducting four-poles in which the magnetic field at the input controls the output resistance), persistors and persistrons (superconducting memory elements for memory devices), etc.

Of highest practical importance is the problem of high temperature superconductivity. Of all the known materials the highest transition temperature is that of the alloy $(\text{Nb}_3\text{Al})_4^+\text{Nb}_3\text{Ge}$ whose $T_{\text{cr}} \approx 20 \text{ K}$. To obtain such a temperature liquid helium is needed. What are the prospects for developing materials with higher critical temperatures?

The BCS theory demonstrates that T_{cr} is directly related to the force of attraction between the electrons in the superconductor and is determined from the following approximate expression:

$$T_{\text{cr}} = \Theta e^{-1/g} \quad (6.61)$$

where Θ is the Debye temperature, and g is a constant (not exceeding $1/2$ and usually less) dependent on the attraction force between the electrons. For $g = 1/3$, the maximum critical temperature obtainable with a material with $\Theta = 500 \text{ K}$ would be $T_{\text{cr}} = \Theta e^{-3} = 0.05\Theta = 25 \text{ K}$. Naturally, this estimate is a very rough one but still it makes it clear that it is impossible to obtain high temperature superconductivity ($T_{\text{cr}} > 100 \text{ K}$) with the electron pairing mechanism discussed above.

Simultaneously, with the development of superconducting materials with increased T_{cr} utilizing the electron pairing effect via positively charged lattice ions, an intensive search for other mechanisms of electronic interaction capable of more efficient attraction and, consequently, of providing superconductive materials with substantially greater transition temperatures T_{cr} , goes on in the laboratories around the world. Should this search be successful and should such materials be produced, the importance of this discovery would prove to be comparable to the development of controlled thermonuclear fusion.

Chapter 7

Magnetic Properties of Solids

§ 63. Magnetic field in magnetic materials

Let us place a homogeneous body of volume V into a uniform magnetic field of intensity \mathbf{H} and induction $\mathbf{B}_0 = \mu_0 \mathbf{H}$. Acted upon by the field, the body becomes magnetized obtaining a magnetic moment M . The ratio of the magnetic moment to the volume of the body is termed *magnetization*, J_m :

$$J_m = \frac{M}{V}, \quad (7.1)$$

and when magnetization is not uniform it is equal to

$$J_m = \frac{dM}{dV}. \quad (7.2)$$

Magnetization is a vector; in uniform magnetic, bodies J_m is either directed parallel or antiparallel to \mathbf{H} . The unit of magnetic moment in the SI system is A m^2 and that of magnetization is A m^{-1} .

The ratio of magnetization J_m to the magnetic field intensity H is termed the *magnetic susceptibility*, χ :

$$\chi = \frac{J_m}{H}. \quad (7.3)$$

It may easily be seen that χ is a dimensionless quantity. From Eq. (7.3), we get

$$J_m = \chi H. \quad (7.4)$$

A magnetized body placed in an external field, establishes its own field which in isotropic magnetic materials, away from its external boundaries, is directed either parallel or antiparallel to the external field. Denote the external field induction by B_0 , the proper field induction by B_1 and the resultant induction by B . For uniform magnetic materials, B is an algebraic sum of B_0 and B_1 :

$$B = B_0 + B_1. \quad (7.5)$$

Experiments show that

$$B_1 = \mu_0 J_m = \chi B_0 \tag{7.6}$$

wherefore

$$B = (1 + \chi) B_0. \tag{7.7}$$

The quantity

$$\mu = 1 + \chi \tag{7.8}$$

is termed *magnetic permeability*. It follows from Eq. (7.8) that

$$\chi = \mu - 1. \tag{7.9}$$

Substituting Eq. (7.8) into (7.7), we obtain

$$B = \mu B_0 = \mu \mu_0 H. \tag{7.10}$$

The unit of field intensity H in the SI system is $A\,m^{-1}$ and that of induction B the tesla (T).

§ 64. Magnetic properties of solids

All materials may be divided into three large groups according to the absolute value and the sign of their magnetic susceptibility (Table 7.1): diamagnetics, paramagnetics and ferromagnetics.

Diamagnetics and paramagnetics. For diamagnetics ($|\chi| < 1$), χ is negative and independent of the intensity of the external magnetic field and of temperature. Such materials are magnetized in the direction opposite to the direction of the external field and because of that they are pushed out of the regions of the highest field intensity.

Paramagnetics also have $|\chi| < 1$, but contrary to diamagnetics χ is positive. Such bodies are magnetized in the direction of the external field and are drawn

Table 7.1

Diamagnetics		Paramagnetics		Ferromagnetics	
Substance	$\chi = \mu - 1$	Substance	$\chi = \mu - 1$	Substance	$\chi = \mu - 1$
Bi	-18×10^{-5}	CaO	580×10^{-5}	Fe	1000
Cu	-0.9×10^{-5}	FeCl ₂	360×10^{-5}	Co	240
Ge	-0.8×10^{-5}	NiSO ₄	120×10^{-5}	Ni	150
Si	-0.3×10^{-5}	Pt	26×10^{-5}		

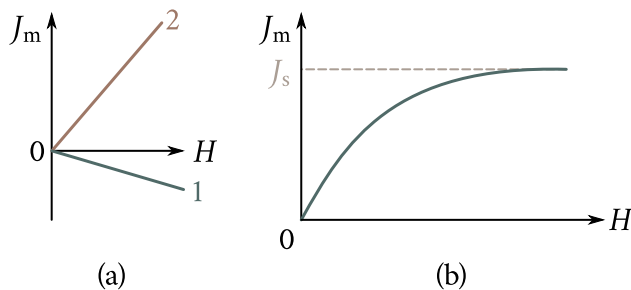


Figure 7.1: Magnetization J_m versus the magnetic field intensity H : (a)—diamagnetics (1) and paramagnetics (2) in weak and medium fields at normal and high temperatures; (b)—paramagnetics at low temperatures (or in very strong fields).

into the regions of maximum H .

Figure 7.1(a) shows the dependence of J_m on the field intensity for diamagnetics, 1, and for paramagnetics, 2. In both cases J_m is proportional to H , this being an indication of the independence of χ of H . However, for paramagnetics this is observed only in relatively weak fields at high temperatures; in strong fields and at low temperatures the plot $J_m(H)$ asymptotically approaches the limit value J_s , which corresponds to magnetic “saturation” of the paramagnetic [Figure 7.1(b)]. Besides, χ of paramagnetic bodies is dependent on temperature. This dependence was first studied by Pierre Curie. He demonstrated that

$$\chi = \frac{C}{T} \quad (7.11)$$

where T is the absolute temperature of the paramagnetic, and C is a constant dependent on its nature. The term for it is the *Curie constant* and for expression (7.11) the *Curie law*.

Ferromagnetics. Magnetic susceptibility χ of ferromagnetic materials, a typical representative of which is iron, is also positive but immeasurably greater than that of paramagnetics. Besides, χ depends on H . Apart from iron this group includes also nickel, cobalt, gadolinium, dysprosium, holmium, erbium and some alloys.

The rules governing the magnetization were first investigated by the Russian physicist A. G. Stoletov. Figure 7.2 shows the dependence of B (a), of magnetization J_m (b), and of susceptibility χ (c) on H for soft iron. B and J_m at first rise quickly with the magnetizing field but then the rise slows down and, at some H_s , a close to the maximum value of J_s is attained; any further slow increase in induction is due solely to the increase in H . This state corresponds to technical saturation of the ferromagnetic: as this state is approached, $\chi \rightarrow 0$.

A careful study of the magnetization curve shows that as H increases J_m rises

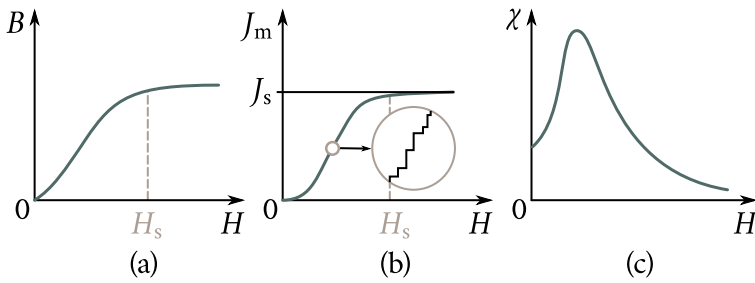


Figure 7.2: Magnetization of ferromagnetics: (a)—induction B versus field intensity H ; (b)—magnetization J_m versus field intensity H (the right-hand side shows a magnified section of the magnetization curve); (c)—magnetic susceptibility versus field intensity H .

not continuously but stepwise. This is especially apparent in the region of the steep rise of the magnetization curve. Figure 7.2(b) shows a magnified section of the magnetization curve (enclosed in a circle). This section consists of a large number of steps corresponding to individual jumps accompanying the variation of J_m with a continuous rise in H . The stepwise nature of the magnetization process was discovered by Heinrich Barkhausen and became known as the *Barkhausen effect*.

Figure 7.3 shows the plot of a complete remagnetizing cycle of a ferromagnetic. It may be seen from Figure 7.3 that during the remagnetization, the variation of B lags behind the variation of H , and when $H = 0$ is not equal to zero but to B_{res} . This lagging of B behind H has been named *magnetic hysteresis* and the induction B_{res} *residual magnetic induction*, or *remanence*. To remove it, a demagnetizing field H_c termed *coersive force* should be applied. The closed curve $AB_{\text{res}}H_cA'B'_{\text{res}}H'_cA$, which describes the remagnetizing cycle, is termed the *hysteresis loop*. The area of this loop is proportional to the work that should be expended to remagnetize a ferromagnetic of unit volume. In the course of remagnetization, this work is completely transformed into heat. Therefore, when the ferromagnetic is remagnetized many times in succession its temperature rises, the effect being the greater the greater the area of the hysteresis loop.

Ferromagnetic materials are classified as “soft” and “hard”, or having a high coersive force. Soft magnetic materials used for manufacturing cores of electric motors and instruments have a low coersive force and high permeability. The best alloys of this type (Supermalloy, for instance) have μ 's as high as 10^6 , saturation induction $B_s \approx 1 \text{ T}$ and a coersive force H_c of only 0.32 A m^{-1} . Their hysteresis-loop area is so small that their remagnetization losses are some 500 times less than those of soft iron. Hard magnetic materials are characterized by a high coersive force and by high residual magnetization. For instance, Magnico, used for manufacturing permanent magnets, has $H_c \approx 5 \times 10^5 \text{ A m}^{-1}$ and $B_{\text{res}} = 1.35 \text{ T}$.

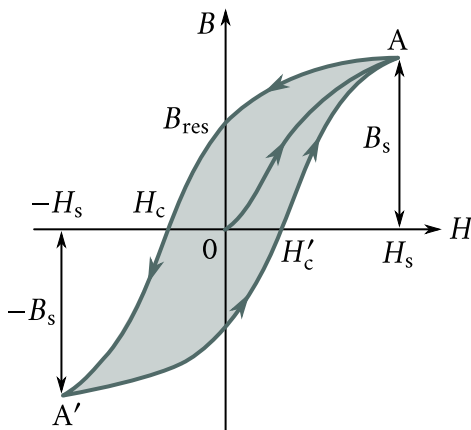


Figure 7.3: Hysteresis loop.

When ferromagnetic materials are heated, their magnetic properties become less pronounced: there is a drop in the values of χ , μ , J_m , etc. There is a temperature Θ_C for every ferromagnetic at which it loses its ferromagnetic properties. This is known as the *ferromagnetic Curie point*. By way of an example we shall show the Curie points of some ferromagnetics: cobalt, 1150 °C; iron, 770 °C; Nickel, 360 °C; 30% permalloy, 70 °C.

Above Θ_C , ferromagnetics turn into paramagnetics with their characteristic linear dependence of $1/\chi$ on T (Figure 7.4), which is quite well represented by the following relation, known as the *Curie-Weiss law*:

$$\chi = \frac{C}{T - \Theta} \quad (7.12)$$

with C the Curie constant and Θ the *paramagnetic Curie point* (it is somewhat higher than Θ_C).

Figure 7.5 shows the temperature dependence of maximum magnetization of iron, nickel, and cobalt. The ratio T/Θ_C is plotted along the x axis and the ratio $J_s(T)/J_s(0)$ along the y axis. In such relative coordinates the dependence of magnetization on temperature is described by the same curve for all ferromagnetics. As temperature rises, magnetization drops becoming practically zero at the Curie point.

Ferromagnetic single crystals are characterized by anisotropic magnetization. Figure 7.6 shows the magnetization curves of iron (a) and nickel (b) crystals in the [111], [110], and [100] directions. It follows from Figure 7.6, that there are directions in the single crystal, in which it is easier to magnetize the crystal and obtain magnetic saturation at relatively small values of magnetic field intensity. Those di-

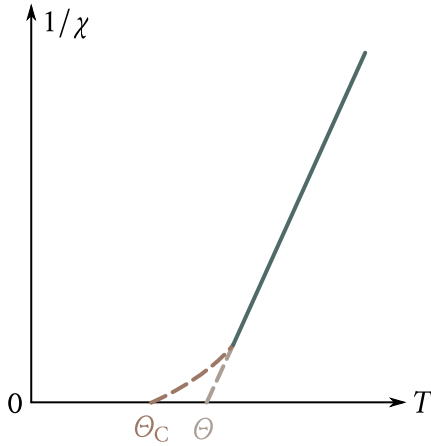


Figure 7.4: Temperature dependence of magnetic susceptibility of ferromagnetics: Θ_C the ferromagnetic Curie point; and Θ the paramagnetic Curie point.

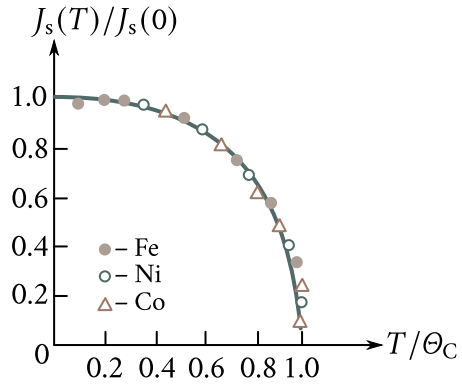


Figure 7.5: Temperature dependence of maximum magnetization of iron, nickel and cobalt.

rections are termed *directions of easy magnetization*. For iron this direction is the $[100]$, and for nickel the $[111]$ direction. It is much more difficult to magnetize iron in the $[110]$ and $[111]$ directions, and nickel in the $[110]$ and $[100]$ directions, substantially greater values of magnetic field intensity being needed to attain magnetic saturation. Those directions are termed *difficult magnetization directions*. The integral

$$W_m = \int_0^{J_s} \mu_0 H dJ_m \quad (7.13)$$

taken along the magnetization curve expresses the work spent on magnetizing the crystal in the given direction. This work is transformed into free energy of the magnetized crystal. It may be seen from Figure 7.6, that the least free energy is that of the crystal magnetized in the easy direction and the greatest is that of the crystal magnetized in the difficult direction.

Magnetization of ferromagnetics is accompanied by a change in their dimensions and shape. This phenomenon became known as *magnetostriction*. Figure 7.7 shows the relative change in the length of rods made of nickel, of annealed and of cast cobalt, of iron, and of steel magnetized in fields of gradually increasing intensity. The greatest relative contraction is that of nickel (almost 0.004 percent); iron and steel rods increase their length a little in weak fields and contract in strong fields. On the contrary, cast cobalt rods contract in weak fields and increase their length in strong fields.

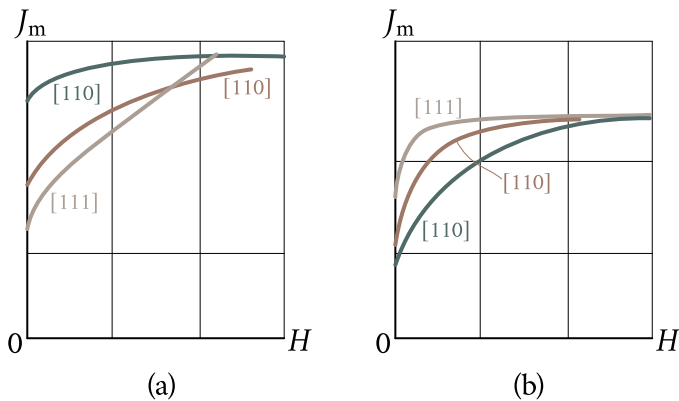


Figure 7.6: Magnetization plots of iron (a) and nickel (b) single crystals in directions [100], [110] and [111].

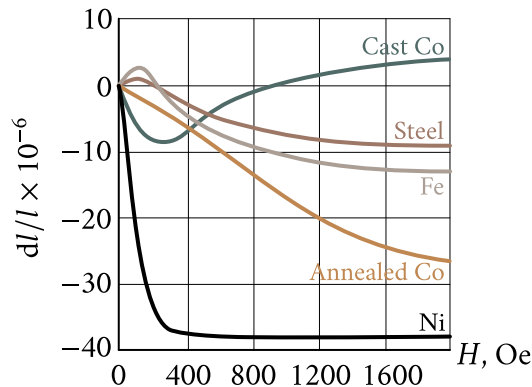


Figure 7.7: Variation of length of ferromagnetic samples with magnetization (magnetostriction).

In compliance with the Le Chatelier principle to the effect that a system resists the influence of external factors striving to change its state, the mechanical deformation of ferromagnetic bodies resulting in the change in their shape and dimensions should influence the magnetization of such bodies. Specifically, if the body being magnetized contracts in the given direction, then application of a compressive stress in this direction should favour magnetization and application of an extending stress should make magnetization more difficult. The variation of magnetic properties of strained ferromagnetic bodies is termed the *magnetoelastic effect*. Some ferromagnetic materials are so sensitive to internal stresses caused by deformations that this property is utilized for the purposes of strain measurements.

When a ferromagnetic is magnetized in an alternating magnetic field, its di-

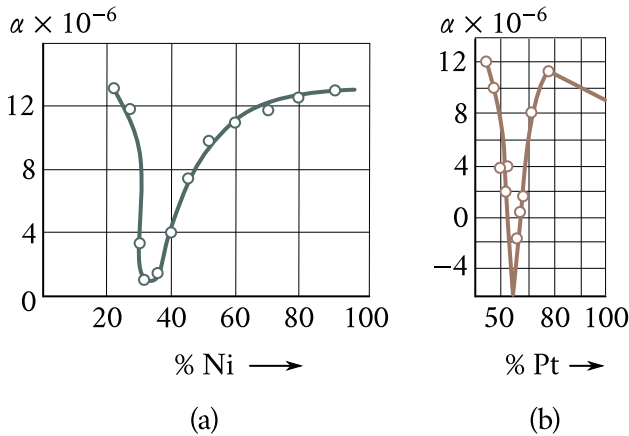


Figure 7.8: Dependence of linear expansion coefficient of iron-nickel (a) and of iron-platinum (b) alloys on composition.

mensions change with a frequency double that of the field. This property is used in *magnetostrictive oscillator* capable of generating powerful ultrasonic vibrations with a frequency up to several megahertz. Such oscillators are employed in ultrasonic devices for the machining and cleaning of solid objects, in sonars used to measure depth of waterways, and in numerous other devices and instruments.

An interesting problem is that of thermal expansion of ferromagnetic bodies. Thermal expansion of solids is, as we know, due to the anharmonicity of vibrations of particles around their equilibrium sites. For diamagnetic and paramagnetic solids, anharmonicity is the only cause of the change in their dimensions upon heating. By force of this such bodies always expand with the rise in temperature. Let us denote the linear expansion coefficient due to anharmonicity of atomic vibrations by α_1 . The situation in ferromagnetic materials is not so simple. A change in their temperature is accompanied by a change in their magnetization and, consequently, in dimensions. N. S. Akulov has proposed the term *thermostriction* for this phenomenon. Denote the linear expansion coefficient due to thermostriction by α_2 . The total thermal expansion coefficient of a ferromagnetic will be $\alpha = \alpha_1 + \alpha_2$. The coefficient α_1 is always positive while α_2 may be either positive or negative. Therefore, the total thermal expansion coefficient of a ferromagnetic material may be positive, zero, or negative. For instance, the group of ferromagnetic materials with a negative “ferromagnetic” component of the thermal linear expansion coefficient includes Invar alloys. Figure 7.8 shows the dependence of the thermal expansion coefficient of iron-nickel (a) and iron-platinum (b) alloys on their composition. The α of alloys containing about 36% nickel is about 10 times

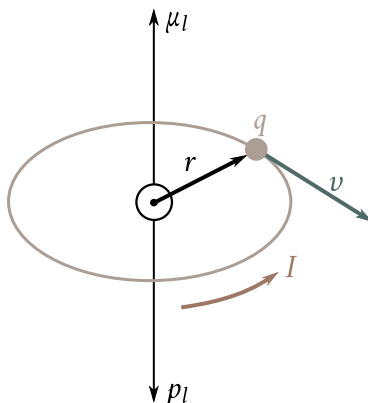


Figure 7.9: Orbital magnetic moment μ_l , and orbital angular momentum p_l of an electron.

less than that of pure nickel or iron: α of alloy containing 56% platinum is negative, such an alloy does not expand upon heating but, on the contrary, contracts.

Invar alloys are widely used in instrument manufacture, metrology, aviation, and manufacture of electric lamps and radio valves. Depending on practical purposes, alloys with very small, zero, or even negative thermal expansion coefficients can be used.

§ 65. Magnetic properties of atoms

The orbital magnetic moment of an atom. The atom of every element is made up of a positively charged nucleus and an electron shell. Many magnetic phenomena can be adequately explained with the aid of Bohr's theory, in which it is assumed that the electrons of the shell move in definite orbits. Each such electron will establish a closed current equal to $I = -q\nu$ (ν is the frequency of rotation of the electron in the orbit, and q its charge). The magnetic moment of the current is $M = IS = -q\nu S$ (S is the area of the orbit). Since $S = \pi r^2$ and $\nu = v/(2\pi r)$ (v is the linear velocity of the electron in the orbit), it follows that

$$M = \mu_l = -\frac{vqr}{2}. \quad (7.14)$$

The magnetic moment of the electron, which is due to its motion around the nucleus, is termed *orbital magnetic moment*. We shall denote it by μ_l . This moment is perpendicular to the plane of the orbit as is required by the right-hand screw rule (Figure 7.9).

The orbital angular momentum of the electron is

$$p_l = mvr \quad (7.15)$$

where m is the electron mass. It is opposite to μ_l . Comparing Eqs. (7.14) and (7.15), we find

$$\mu_l = -\frac{q}{2m} p_l. \quad (7.16)$$

The ratio

$$\gamma_l = \frac{\mu_l}{p_l} = -\frac{q}{2m} \quad (7.17)$$

is termed the *gyromagnetic ratio*.

As required by quantum mechanics, \mathbf{p}_l and its projection p_{lH} on the direction of the magnetic field H , can assume only discrete values, namely

$$p_l = \hbar[l(l+1)]^{1/2}, \quad (7.18)$$

$$p_{lH} = m_l \hbar, \quad (7.19)$$

where l is the orbital quantum number, which can assume only the following values:

$$l = 0, 1, 2, \dots, n \quad (7.18')$$

n values in all (n is the principal quantum number); m_l is the magnetic quantum number, which can assume only the following values:

$$m_l = -l, -(l-1), \dots, 0, \dots, +l \quad (7.19')$$

$2l+1$ values in all.

Because of that, the magnetic moment μ_l and its projection μ_{lH} on the direction of H may assume only the following discrete values:

$$\mu_l = -\frac{q\hbar}{2m} [l(l+1)]^{1/2} = -\mu_B [l(l+1)]^{1/2}, \quad (7.20)$$

$$\mu_{lH} = -m_l \mu_B, \quad (7.21)$$

where

$$\mu_B = -\frac{q\hbar}{2m} = 9.27 \times 10^{-24} \text{ A m}^2 \quad (7.22)$$

is the *Bohr magneton*. It is the “quantum” of the magnetic moment and is accepted as a unit for measuring magnetic moments of atomic systems.

In a complex atom whose electron shell is made up of many electrons, the total orbital magnetic moment is found by adding up the moments of individual electrons in compliance with the rules of space quantization. The moment of closed electron shells is zero. By force of this only, the atoms with partially filled shells can have a nonzero orbital magnetic moment. But even in the latter case, should the partially filled shell lie close to the external shell and should the interaction of the atoms in the solid state be strong, the magnetic moments of the partially filled shell would be “frozen in”: their orientation in an external field would be so im-

paired that they would take practically no part in the magnetization of the body. For instance, such is the behaviour of orbital magnetic moments of the electrons of the partially filled 3d shell of the elements belonging to the iron group.

The spin magnetic moment of an atom. Apart from the orbital angular momentum, the electron has an intrinsic angular momentum p_s termed *spin*. It is known from quantum mechanics that

$$p_s = \sqrt{3} \frac{\hbar}{2} \quad (7.23)$$

and that the projection of the spin on the direction of the field H may assume only the following values:

$$p_{sH} = \pm \frac{\hbar}{2}. \quad (7.24)$$

There is an intrinsic magnetic moment μ_s connected with the intrinsic electron angular momentum whose value was first experimentally determined by Otto Stern and Walther Gerlach. Their experiments demonstrated that the projection μ_{sH} is numerically equal to the Bohr magneton:

$$\mu_{sH} = \pm \mu_B = \pm \frac{q\hbar}{2m} = -\frac{q}{m} p_{sH} \quad (7.25)$$

(the minus sign reflects the negative nature of the electron charge). The gyromagnetic ratio for the intrinsic moments of the electron is

$$\gamma_s = \frac{\mu_{sH}}{p_{sH}} = -\frac{q}{m}. \quad (7.26)$$

It is twice as large as γ_l for the orbital moments.

In atoms containing a large number of electrons, p_s should be added up like vectors with account taken of the rules of space quantization. The total spin moment of closed shells is zero, the same as the orbital moment. Table 7.2 shows, by way of an example, the data on the spin configuration of the 3d shell of free atoms of the elements of the iron group.

The spins are least compensated in the chromium and manganese atoms and, correspondingly, they have the maximum total spin moment. However, such ori-

Table 7.2

	Sc	Ti	V	Cr	Mn	Fe	Co	Ni
Total spin	1 ↓	2 ↓↓	3 ↓↓↓	5 ↓↓↓↓↓	5 ↓↓↓↓↓	4 ↑↓↓↓↓	3 ↑↑↑↓↓	2 ↑↓↑↑↓↓

entation of the spins is not usually retained in the solid state and, because of that, the total atomic spin moment in the solid is different. For instance, in the iron lattice, the average number of Bohr magnetons per atom is not 4 but only 2.3; in chromium it is 0.4, and in α -manganese it is 0.5.

Magnetic moments of nucleus. Atomic nuclei too have a spin and a magnetic moment connected with it. The order of magnitude of the nuclear spin is the same as that of the electron. Since the nuclear mass is some 10^3 times greater than the electron mass, the nuclear magnetic moment, in compliance with Eq. (7.25), is three orders of magnitude less than the electron magnetic moment. Therefore, as a first approximation, the effect of nuclear magnetic moments on the magnetic properties of bodies can be neglected. This does not mean that those moments do not play any role at all. In some phenomena, not discussed in this book, that role may be quite important.

The total magnetic moment of an atom. The total magnetic moment of the electron shell of the atom is determined as follows. Using the rules of space quantization, one finds the total orbital angular momentum: $P_L = \sum_i p_{li}$, where p_{li} is the orbital angular momentum of the i -th electron. The numerical value of P_L is determined by the quantum number L

$$P_L = \hbar[L(L+1)]^{1/2}. \quad (7.27)$$

The number L may be any integer between the maximum and the minimum values of the algebraic sum $\sum_i l_i$ of the orbital quantum numbers l_i of individual electrons. Next, one finds the total atomic spin: $P_S = \sum_i p_{si}$, where p_{si} is the spin of the i -th electron. The numerical value of P_S is determined by the quantum number S :

$$P_S = \hbar[S(S+1)]^{1/2}. \quad (7.28)$$

The number S may assume values lying in the interval between the maximum and the minimum values of the algebraic sum $\sum_i s_i$ of spin quantum numbers of the individual electrons, the difference between successive values of S being unity.

Finally, one finds the total atomic momentum \mathbf{P}_J , equal to the vector sum of \mathbf{P}_L and \mathbf{P}_S , that is $\mathbf{P}_J = \mathbf{P}_L + \mathbf{P}_S$. The numerical value of \mathbf{P}_J is determined by the intrinsic quantum number J :

$$P_J = \hbar[J(J+1)]^{1/2}. \quad (7.29)$$

which may assume the following set of values:

$$\begin{aligned} J &= L + S, L + S - 1, \dots, L - S, & \text{if } L > S \\ J &= S + L, S + L - 1, \dots, S - L, & \text{if } S > L. \end{aligned} \quad (7.30)$$

The only allowed orientations of \mathbf{P}_J in an external field are such that its projections

on the direction of the field are multiples of \hbar :

$$P_{JH} = m_J \hbar, \quad (7.31)$$

where m_J is the magnetic quantum number equal to

$$m_J = -J, -(J-1), \dots, 0, 1, 2, \dots, J \quad (7.32)$$

$2J+1$ values in all.

The atomic magnetic moment corresponding to the total momentum P_J is

$$M_J = -g\mu_B [J(J+1)]^{1/2} \quad (7.33)$$

with projections on the direction of an external field H equal to

$$M_{JH} = -m_J g \mu_B \quad (7.34)$$

where

$$g = 1 + \frac{J(J+1) + S(S+1) - L(L+1)}{2J(J+1)} \quad (7.35)$$

is the *Lande factor* or *magnetic splitting factor*, which takes account of the difference in gyromagnetic ratios of the orbital and the spin moments making up the total atomic magnetic moment. For $L = 0$, that is, in the case of a purely spin magnetism, $g = 2$; for $S = 0$, that is, in the case of a purely orbital magnetism, $g = 1$.

Often the term atomic magnetic moment is taken to mean not Eq. (7.33), but the maximum value of the projection M_{JH} . For instance, the magnetic moment of a hydrogen atom in the ground state ($L = 0, S = 1/2, g = 2$) characterized by $J = 1/2$ is taken to be equal to μ_B ; for a free iron atom with a “frozen in” orbital magnetic moment, $J = S = 2, g = 2$ and $M_a = 4\mu_B$.

All atoms and ions with closed shells have $S = 0, L = 0$, and $J = 0$. Therefore, the magnetic moments of such atoms and ions are zero. Paramagnetism owes its existence to the presence in an atom of partially filled shells. According to the Pauli exclusion principle, there may not be more than two electrons with opposite spins in one state. The total spin moment of those electrons is zero. Such electrons are termed *paired*. If an atom or an ion contains an odd number of electrons, one of them will be unpaired and the atom will have a permanent magnetic moment. If the atom contains an even number of electrons, two cases are possible: either all electrons are paired and the total spin moment is zero or two or more electrons are unpaired and the atom has a permanent magnetic moment. For instance, H, K, Na, Ag have odd numbers of electrons, one of them unpaired; Be, C, He, Mg contain even numbers of electrons, all of them paired. Oxygen also contains an even number of electrons, but two of them are unpaired.

Magnetic moments of many molecules are zero because only some of them contain unpaired electrons. First of all, these are the free radicals, which play an exceptionally important part in many chemical reactions. As examples of such, rad-

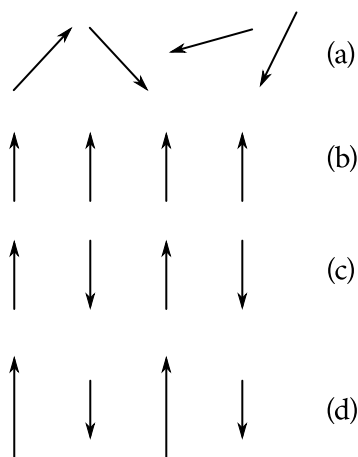


Figure 7.10: Schematic representation of atomic magnetic moments in paramagnetic (a), ferromagnetic (b), antiferromagnetic (c), and ferrimagnetic (d) materials.

icals are free hydroxyl (OH), free methyl (CH_3), and free ethyl (C_2H_5). The presence of unpaired electrons in molecules and in free radicals makes them magnetic.

Classification of magnetic materials. When the orbital and the spin moments are added up, a complete compensation may take place and then the total atomic moment will be zero. If such a compensation does not take place, the atom will have a permanent magnetic moment. Accordingly, the magnetic properties of bodies will be different.

Materials whose atoms have no permanent magnetic moments are diamagnetic. Materials whose atoms have a permanent magnetic moment may be either paramagnetic, ferromagnetic, antiferromagnetic, or ferrimagnetic. Namely, if the interaction between the atomic magnetic moments is zero or very weak, the material will be paramagnetic [Figure 7.10(a)]; if the neighbouring magnetic moments tend to align themselves parallel to one another, the material will be ferromagnetic [Figure 7.10(b)]; if the neighbouring magnetic moments tend to align themselves antiparallel to one another, the material will be antiferromagnetic [Figure 7.10(c)]; finally, if the neighbouring magnetic moments tend to align themselves antiparallel to one another but their magnitude is not the same, then the material will be ferrimagnetic [Figure 7.10(d)].

§ 66. Origin of diamagnetism

The cause of diamagnetism is a change in the orbital motion of the electrons acted upon by an external magnetic air field. It is common to materials but is often over-

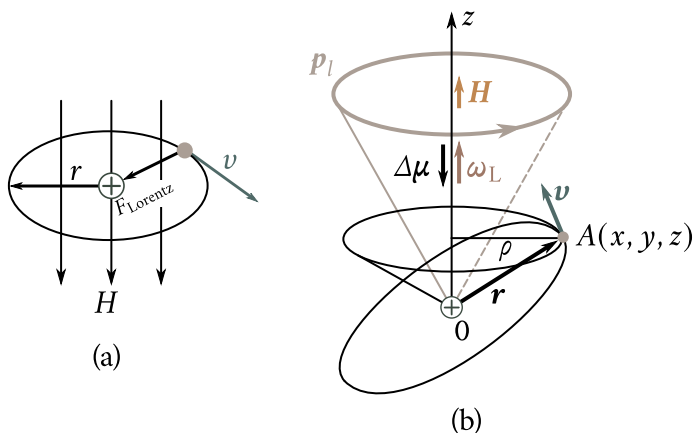


Figure 7.11: Effect of magnetic field on orbital motion of an electron: (a)—field H is perpendicular to orbit plane; (b)—orbit precession in magnetic field.

shadowed by strong para- and ferromagnetism. In its pure form, diamagnetism is displayed by materials whose total atomic magnetic moment is zero.

Precession of electron “orbits” in a magnetic field. Consider the motion of an electron in an orbit of radius r [Figure 7.11(a)]. In the absence of field H , the centripetal force applied to the electron is $F_{cp} = mv_0^2/r = m\omega_0^2 r$ (v_0 is the linear and ω_0 the angular velocity of the electron’s motion). When external field H , perpendicular to the plane of the orbit is applied, the electron is acted upon by the Lorentz force $F_L = qv_0 B_0$ directed along the radius of the orbit (B_0 is the field’s induction). The resultant centripetal force will be

$$F = F_{cp} + F_{Lorentz} \quad \text{or} \quad m\omega^2 r = m\omega_0^2 r + q\omega_0 r B_0.$$

It follows that

$$mr(\omega^2 - \omega_0^2) = mr(\omega - \omega_0)(\omega + \omega_0) \approx 2mr\omega_0\omega_L = q\omega_0 B_0 \quad (7.36)$$

where

$$\omega_L = \omega - \omega_0 = \frac{q}{2m} B_0 = \frac{q}{2m} \mu_0 H \quad (7.37)$$

is called the *Larmor angular frequency*.

Thus, a magnetic field changes the angular frequency of an orbiting electron. It may be seen from Eq. (7.37) that this change is the same for all electrons no matter what the radius of their orbits and the linear velocity of their motion are. The direction of ω_L coincides with that of B_0 .

Generally, when H is not perpendicular to the plane of the orbit, its effect is to excite precession of the orbit around the direction of the field [Figure 7.11(b)]: the perpendicular p_l to the plane of the orbit describes a cone around H . Calculations

show that the angular velocity of such a precession is expressed by Eq. (7.37).

Induced magnetic moment of an atom. Magnetic susceptibility of diamagnetics. The precession of the electron orbit results in an additional motion of the electron around field H . This motion is superimposed on its orbital motion. The magnetic action of this additional motion is equivalent to that of a closed current

$$\Delta I = -qv_L = -q\frac{\omega_L}{2\pi} = -\frac{q^2}{4\pi m}B_0 \quad (7.38)$$

where v_L is the precession frequency related to the angular frequency by the expression

$\omega_L = 2\pi v_L$. The minus appears because of the negative charge of the electron.

The magnetic moment of the elementary current ΔI is

$$\Delta\mu = \Delta IS = -\frac{q^2 S}{4\pi m}B_0 \quad (7.39)$$

where S is the area bounded by the path of the electron precessing around field H . Calculations show that $S = 2\pi\overline{r^2}/3$, where $\overline{r^2}$ is the mean square of the electron's distance from the nucleus. Therefore,

$$\Delta\mu = -\frac{q^2\overline{r^2}}{6m}\mu_0 H. \quad (7.40)$$

It follows from Eq. (7.40) that in a magnetic field every electron acquires an additional so-called *induced magnetic moment* directed against H . The appearance of this moment is the cause of the magnetization of the body in the direction opposite to that of the magnetic field, which is characteristic of diamagnetics.

The magnetic moment of an atom containing Z electrons is found by adding up the moments of individual electrons:

$$\Delta M = -\frac{q^2 B_0}{6m} \sum_i^Z \overline{r_i^2} \quad (7.41)$$

where $\overline{r_i^2}$ is the mean square distance of the i -th electron from the nucleus. The sum of $\overline{r_i^2}$ may be replaced by the product $Z\overline{a^2}$, where $\overline{a^2}$ is the mean square distance of all the electrons from the nucleus. Then,

$$\Delta M = -\frac{Z\overline{a^2}q^2}{6m}B_0. \quad (7.42)$$

Multiplying Eq. (7.42) by the number of atoms per unit volume, n , we obtain the magnetization J_m :

$$J_m = n\Delta M = -\frac{Z\overline{a^2}q^2 n}{6m}B_0 = -\frac{Z\overline{a^2}q^2 n}{6m}\mu_0 H. \quad (7.43)$$

The magnetic susceptibility is

$$\chi = \frac{J_m}{H} = -\frac{\mu_0 Z \bar{a}^2 q^2 n}{6m}. \quad (7.44)$$

Assuming that $a \approx 10^{-10}$ m and $n \approx 5 \times 10^{-28}$ m⁻³, we obtain $\chi \approx 10^{-6} Z$. This is in good agreement with the data of Table 7.1. Moreover, from Eq. (7.44), it follows that magnetic susceptibility of diamagnetics is independent both, of temperature and of magnetic field intensity H , and rises in proportion to the atomic number of the element, Z , which is in full agreement with experiment.

§ 67. Origin of paramagnetism

Langevin's classical theory of paramagnetism. The classical theory of paramagnetism developed by Paul Langevin is based on the idea that the atoms of paramagnetic materials have a permanent magnetic moment \mathbf{M} , that is, they constitute permanent magnetic dipoles and that the interaction between these dipoles is negligible. The energy of such a dipole in a magnetic field \mathbf{H} is

$$U_m = -M\mu_0 H \cos \theta \quad (7.45)$$

where θ is the angle between \mathbf{M} and \mathbf{H} [Figure 7.12(a)].

The minimum of U_m corresponds to $\theta = 0$. Therefore, all the dipoles tend to orient themselves in the direction of the external field, this being hampered by thermal motion. The total magnetic moment of the material is made up of the projections of the magnetic moments of the individual atoms on the direction of \mathbf{H} . Since the magnitude of those projections is $M_H = M \cos \theta$, the problem of the quantitative calculation of the magnetization of the material is reduced to the calculation of the average value of M_H that corresponds to the state of equilibrium between the orientational effect of the field and the disorientational effect of thermal motion. Just this problem was solved by Langevin with the aid of methods of classical statistics. He supposed that the orientation of \mathbf{M} with respect to \mathbf{H} can be arbitrary and, that accordingly the angle θ , can assume all values.

The probability for a dipole to align itself at an angle in the interval $(\theta, \theta + d\theta)$ to \mathbf{H} [that is inside the solid angle $d\Omega$; see Figure 7.12(b)] is determined by the Boltzmann distribution function:

$$W = C_1 e^{-U_m/k_B T} d\Omega = C_1 \exp\left(\frac{\mu_0 M H \cos \theta}{k_B T}\right) d\Omega$$

where C_1 is a normalization constant.

It may be seen from Figure 7.12(b) that $d\Omega = 4\pi \sin \theta d\theta$; therefore,

$$W = C_2 \exp\left(\frac{\mu_0 M H \cos \theta}{k_B T}\right) \sin \theta d\theta$$

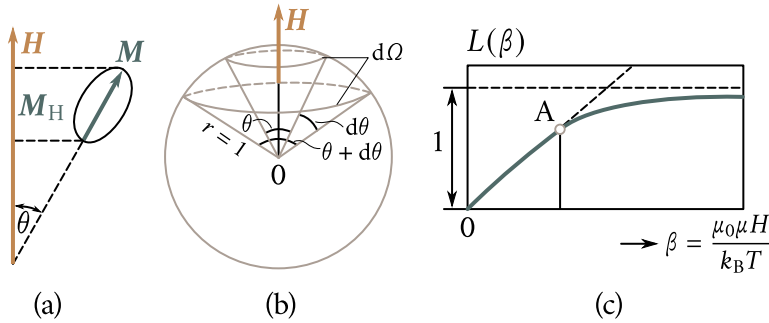


Figure 7.12: Explaining classical theory of paramagnetism: (a)—magnetic moment \mathbf{M} and its projection \mathbf{M}_H on magnetic field \mathbf{H} ; (b)—calculating total magnetic moment of a paramagnetic; (c)—plot of the Langevin function.

where C_2 is a new constant.

The average value of M_H is

$$\overline{M}_H = \overline{M \cos \theta} = M \frac{\int_0^\pi \cos \theta \exp\left(\frac{\mu_0 M H \cos \theta}{k_B T}\right) \sin \theta d\theta}{\int_0^\pi \exp\left(\frac{\mu_0 M H \cos \theta}{k_B T}\right) \sin \theta d\theta}. \quad (7.46)$$

If those integrals are evaluated, the result is

$$\overline{M}_H = M \left[\left(\frac{e^\beta + e^{-\beta}}{e^\beta - e^{-\beta}} \right) - \frac{1}{\beta} \right] = M \left(\coth \beta - \frac{1}{\beta} \right) \quad (7.47)$$

where

$$\beta = \frac{M \mu_0 H}{k_B T}. \quad (7.48)$$

The magnetization is

$$J_m = n \overline{M}_H = n M \left(\coth \beta - \frac{1}{\beta} \right) \quad (7.49)$$

where n is the number of atoms per unit volume, and the magnetic susceptibility is

$$\chi = \frac{J_m}{H} = \frac{n M}{H} \left(\coth \beta - \frac{1}{\beta} \right). \quad (7.50)$$

Since the atomic dipoles acted upon by a field align themselves in its direction, such will be the direction of the magnetization of the body as a whole, which is characteristic of paramagnetics.

Let us expand $\coth \beta$ in a power series: $\coth \beta = \beta^{-1} + \beta/3 - \beta^2/45 + \dots$. For

$\beta \ll 1$, we can limit ourselves with the first two terms of the expansion. Then,

$$J_m = \frac{nM\beta}{3} = \frac{nM^2}{3k_B T} \mu_0 H, \quad \chi = \frac{n\mu_0 M^2}{3k_B T}. \quad (7.51)$$

In full agreement with experiment J_m is directly proportional to H and inversely proportional to T . The second of Eqs. (7.51), expresses the Curie law: $\chi = C/T$. The Curie constant $C = n\mu_0 M^2 / (3k_B)$.

For atoms $M \approx \mu_B$; then, for $H \approx 10^6 \text{ A m}^{-1}$, we see that $MH\mu_0 \approx 10^{-23} \text{ J}$ and for $T = 300 \text{ K}$, we see that $k_B T \approx 3 \times 10^{-21} \text{ J}$. Hence, the condition $\beta \ll 1$ is almost always satisfied. Only in very strong fields and at very low temperatures is $\beta \gg 1$ and the direct proportionality between J_m and H is no longer maintained. In the limiting process, as $\beta \rightarrow \infty$, $\coth \beta \rightarrow 1$, and the magnetization becomes saturated, the corresponding maximum value being

$$J_s = nM. \quad (7.52)$$

Magnetic saturation involves the alignment of magnetic moments of all atoms in the direction of the field.

The function $L(\beta) = \coth \beta - 1/\beta$ is termed the *Langevin junction*. Its plot is shown in Figure 7.12(c). For small β 's, a good approximation for the plot is the segment of the straight line $0A$; as $\beta \rightarrow \infty$, the function $L(\beta) \rightarrow 1$.

Fundamentals of quantum theory of paramagnetism. The classical theory is incapable of providing a consistent explanation of the magnetic phenomena as the result of the motion of electric charges. The existence of molecular currents necessarily involves the acknowledgment of the fact of the stability of electronic motion in atoms, a fact unacceptable for classical physics. The assumption that all orientations of magnetic moments with respect to H are possible, which is the basis of Langevin's classical theory, is also wrong. Those difficulties have, on the whole, been overcome by the quantum theory of paramagnetism. Let's consider briefly the essence of this theory.

There are $2J+1$ ways in which the atomic magnetic moment M_J may align itself in a magnetic field (J is the intrinsic quantum number). The probability of each such orientation is determined by the Boltzmann distribution $W = C e^{\mu_0 M_{JH} H / k_B T}$ (M_{JH} is the projection of M_J on H). The average value of M_{JH} will be

$$\overline{M}_{JH} = \frac{\sum_{-J}^{+J} M_{JH} \exp\left(\frac{\mu_0 M_{JH} H}{k_B T}\right)}{\sum_{-J}^{+J} \exp\left(\frac{\mu_0 M_{JH} H}{k_B T}\right)}. \quad (7.53)$$

The difference between Eq. (7.53) and the classical expression (7.46) is that inte-

gration is replaced by summation over the discrete directions in which the vector \mathbf{M}_J may be aligned. Evaluation of the sums in Eq. (7.53) yields the following result:

$$\overline{M}_{JH} = gJ\mu_B B_J(\beta) \quad (7.54)$$

where

$$\beta = \frac{gJ\mu_B H\mu_0}{2k_B T}, \quad (7.55)$$

$$B_J(\beta) = \frac{2J+1}{2J} \coth\left(\frac{2J+1}{2J}\beta\right) - \frac{1}{2J} \coth\left(\frac{1}{2J}\beta\right). \quad (7.56)$$

Function $B_J(\beta)$ is termed the *Brillouin function*.

The magnetization and the magnetic susceptibility are equal to

$$J_m = n\overline{M}_{JH} = ngJ\mu_B B_J(\beta), \quad (7.57)$$

$$\chi = \frac{ngJ\mu_B}{H} B_J(\beta). \quad (7.58)$$

For $\beta \ll 1$, $B_J(\beta) \approx \beta(J+1)/(3J)$ and

$$J_m = \frac{ng^2\mu_B^2 J(J+1)\mu_0 H}{3k_B T}, \quad \chi = \frac{nJ(J+1)g^2\mu_B^2\mu_0}{3k_B T}. \quad (7.59)$$

It follows from Eq. (7.59) that for $\beta \ll 1$ the quantum theory results in a linear dependence of J_m on H and in an inverse dependence of J_m and χ on T , which agrees with experiment. In strong fields and at very low temperatures $\beta \rightarrow \infty$,

$$\coth\left(\frac{2J+1}{2J}\beta\right) \rightarrow 1, \quad \coth\left(\frac{1}{2J}\beta\right) \rightarrow 1, \quad B_J(\beta) \rightarrow 1$$

and the magnetization attains the saturation value

$$J_s = ngJ\mu_B. \quad (7.60)$$

Materials used for experimental tests of the theory of paramagnetism are the solutions and crystalline hydrates of salts, which contain ions with nonzero magnetic moment. Such are, for instance, the ions of the elements of the groups of iron and rare earths. In solutions and in crystalline hydrates the ions are so far apart, that their interaction may be neglected, which is a necessary condition for paramagnetism. Experimental investigations of such compounds have produced results in good agreement with the theory.

Paramagnetism of electron gas. According to Eqs. (7.51) and (7.59), paramagnetic susceptibility is inversely proportional to temperature. However, some metals have been discovered to exhibit paramagnetism independent of temperature. It was Wolfgang Pauli who demonstrated that this is due to the paramagnetism of free electrons that constitute the electron gas.

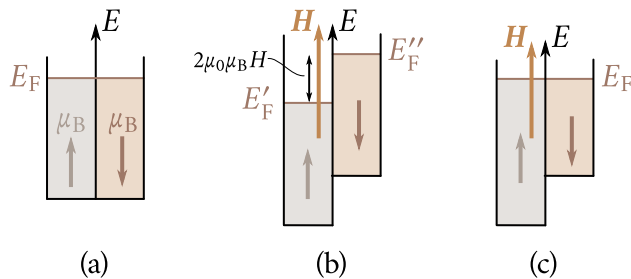


Figure 7.13: Calculating paramagnetism of free electrons.

Figure 7.13(a) shows the conduction band of a metal. It is schematically represented in the form of two half-bands containing electrons with opposite spin moments $\mu_s = \mu_B$. When $H = 0$, the number of electrons in both half-bands is equal and the total magnetic moment of the electron gas is zero. When the field H is applied, every electron of the left half-band acquires an additional energy $U'_m = -\mu_0\mu_B H$ and every electron of the right half-band an energy $U''_m = \mu_0\mu_B H$. The result is the appearance of a difference between the quasi-Fermi levels $E'_F - E''_F = 2\mu_0\mu_B H$ [Figure 7.13(b)] for the electrons of the right and the left half-bands which is equalized by means of spin-flip of some of the electrons of the right half-band and their transition to the left half-band [Figure 7.13(c)]. Since all the internal levels in the half-bands are occupied, the only electrons whose spins can be reversed are those occupying levels in the zone where the Fermi distribution is fuzzy [see Figure 5.6(b)] and where there are vacant levels. The number of such electrons, according to Eq. (3.43), is

$$\Delta n \approx \frac{k_B T}{E_F} n \quad (7.61)$$

where n is the electron gas concentration.

Of this number, $\Delta n' = C \exp [\mu_0\mu_B H / (k_B T)]$ electrons will have their magnetic moments oriented in the direction of H , and $\Delta n'' = C \exp [-\mu_0\mu_B H / (k_B T)]$ against H (C is a constant). The magnetic moment per unit volume of a metal due to spin-flip is

$$J_{me} = (\Delta n' - \Delta n'') \mu_B = C \mu_B (e^\beta - e^{-\beta})$$

where

$$\beta = \frac{\mu_0\mu_B H}{k_B T}.$$

Since $\Delta n = \Delta n' + \Delta n'' = C (e^\beta + e^{-\beta})$, it follows that $C = \Delta n (e^\beta + e^{-\beta})^{-1}$.

Substituting this into the expression for J_{me} , we find

$$J_{me} = \Delta n \mu_B \left(\frac{e^\beta - e^{-\beta}}{e^\beta + e^{-\beta}} \right) = \Delta n \mu_B \tanh \beta.$$

For $\beta \ll 1$, $\tanh \beta \approx \beta$ and $J_{me} = \Delta n \mu_B^2 \mu_0 H / (k_B T)$. Substituting Δn from Eq. (7.61), we obtain

$$J_{me} \approx n \frac{\mu_B^2}{E_F} \mu_0 H. \quad (7.62)$$

The paramagnetic susceptibility of the electron gas is

$$\chi_e \approx n \frac{\mu_0 \mu_B^2}{E_F}. \quad (7.63)$$

A more accurate calculation yields

$$\chi_e = \frac{3}{2} n \frac{\mu_0 \mu_B^2}{E_F}. \quad (7.64)$$

It may be seen from Eq. (7.64) that the magnetic susceptibility of the electron gas should be independent of temperature, which is what is observed in practice.

Production of low temperatures using the method of adiabatic demagnetization of paramagnetic samples. The atoms of paramagnetic materials possess a permanent magnetic moment. In the absence of an external field, as a result of thermal motion of the atoms, the orientation of their magnetic moments is almost completely random. The quantitative measure of this disorder is the entropy S , which in this case is termed magnetic entropy S_M . In compliance with the Boltzmann principle

$$S_M = k_B \ln W_M \quad (7.65)$$

where W_M is the thermodynamic probability, which in this case, is equal to the number of ways the n atoms of the paramagnetic sample can be distributed among the $2J + 1$ sublevels into which every atomic level splits in a magnetic field. Its value may be obtained from the expression

$$W_M = (2J + 1)^n. \quad (7.66)$$

Substituting Eq. (7.66) into (7.65) we obtain

$$S_M = k_B n \ln (2J + 1). \quad (7.67)$$

When the magnetic field is applied and its intensity increased, an ever increasing number of magnetic moments is oriented in the direction of the field, the result being a reduction in the magnetic entropy. When the state of magnetic saturation is reached, the greatest possible order in the arrangement of magnetic moments is established and S_M vanishes. Hence, the process of magnetization of a paramagnetic sample up to saturation is accompanied by the decrease in its entropy by the

amount

$$\Delta S = S_M = k_B n \ln (2J + 1). \quad (7.68)$$

If the magnetization is performed at a constant temperature T , the decrease in S by the amount ΔS results in the generation of an amount of heat equal to $\Delta Q = T \Delta S$. This heat is transmitted from the sample to the surroundings, usually to liquid helium. After equilibrium has been established, the helium is removed and the sample is left thermally insulated. In such conditions it is slowly adiabatically demagnetized with the result that its entropy again rises by ΔS . The rise in entropy requires heat, which can be supplied only by the thermal vibrations of the lattice, since the sample is thermally insulated from the surroundings. Because of that, its temperature drops. Using this method it was possible to obtain temperatures below 0.001 K. The possibility of obtaining still lower temperatures is limited mainly by the fact that already at $H = 0$, the atomic energy levels are, to some extent, split into sublevels because of the interaction of the magnetic moments with each other and with the nucleus.

§ 68. Origin of ferromagnetism

Elementary carriers of ferromagnetism. A magnetized body acquires a magnetic moment M made up of regularly oriented atomic magnetic moments and an angular momentum P made up of regularly oriented atomic angular momenta. According to Eqs. (7.17) and (7.26), the ratio M/P must be equal to $q/2m$ if the magnetization is due to orbital magnetic moments of the atoms, and to q/m if it is due to spin moments.

The appearance of a magnetic moment in the course of magnetization was first established in experiments of A. Einstein and W. J. de Haas and became known as the *Einstein-de Haas effect*. In those experiments, a small iron rod 1 suspended on a thin elastic thread 2 was placed inside a solenoid 3 [Figure 7.14(a)]. In the course of magnetization the rod turned and twisted the thread. The direction of rotation of the rod changed with the change in the direction of magnetization. The angle of rotation was measured with the aid of mirror 4 fixed on the rod which reflected a beam of light on scale 5. The experiment made it possible to determine M and P and to find the gyromagnetic ratio $\gamma = M/P$.

S. J. Barnett made an experiment which was the reverse of the Einstein-de Haas experiment: he observed the magnetization of a quickly rotating iron rod. Such magnetization is caused by the tendency of electrons—thought of as tops possessing angular momenta—to arrange their axes of rotation (spins) in the direction of the rotation axis of the body [Figure 7.14(b)]. In another experiment, the Soviet

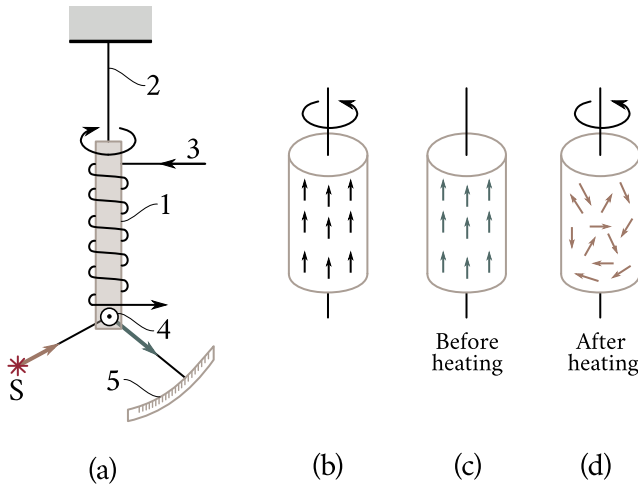


Figure 7.14: Experiments on the nature of ferromagnetism: (a)—the experiment of Einstein and de Haas; (b)—experiment of Barnett; (c,d)—experiment of Ioffe and Kapitza.

physicists A. F. Ioffe and P. L. Kapitza quickly heated a magnetized rod to a temperature above the Curie point. Before heating, the orientation of the “electron tops” was ordered [Figure 7.14(c)] and their total angular momentum was nonzero. When heated to a temperature above the Curie point the “tops” changed their orientation to chaotic [Figure 7.14(d)] and their total angular momentum became zero. Because of that, the demagnetized rod as a whole acquired a rotational momentum that could be measured in the experiment. Measuring in addition the magnetic moment of the magnetized rod one could find the gyromagnetic ratio $\gamma = M/P$.

The experiments demonstrated that the gyromagnetic ratio for ferromagnetic materials is $M/P = -q/m$, that is, equal to the gyromagnetic ratio for the intrinsic moments of the electron. This proved that ferromagnetism is due not to orbital but to spin magnetic moments of electrons, which is consistent with the electronic structure of the atoms of elements that exhibit ferromagnetism. Since the magnetic moments of closed shells are zero and since the outer valence electrons are collectivized in the process of formation of the metallic state, ferromagnetism must be a property solely of the transitional elements, which have incomplete inner shells. Those elements include the transitional metals of the iron group, which have an incomplete 3d shell, and rare earth elements with an incomplete 4f shell. Since, on the other hand, the orbital magnetic moments of the electrons of the 3d shell are “frozen in” and their contribution to the magnetic properties is negligible, ferromagnetism of the elements belonging to those groups can be due only to the atomic spin moments, which in this case are quite large (Table 7.2). This hypothesis was

first expressed by the Russian scientist B. Rozing in 1892 and was exploited later by the French physicist Pierre Weiss. The latter assumed that there is an intense molecular field H in a ferromagnetic proportional to the saturation magnetization J_s :

$$H = \lambda J_s \quad (7.69)$$

where λ is termed the *molecular field constant*. This field is responsible for the spontaneous magnetization of a ferromagnetic.

The introduction of a molecular field made it possible to explain a wide range of phenomena observed in ferromagnetism. However, the nature of the field itself remained a mystery for a long time. At first, it was supposed that the origin of forces which orient the spin moments is purely magnetic, that they appear as a result of an ordinary interaction of spin magnetic moments (spin-spin interaction). The energy of this interaction is of the order of $U_m \approx \mu_B^2/a^2$, where a is the interatomic distance in the lattice of the ferromagnetic. Substituting $\mu_B = 9.27 \times 10^{-24} \text{ A m}^2$ and $a \approx 10^{-10} \text{ m}$, we obtain $U_m \approx 10^{-23} \text{ J}$. This is about two orders of magnitude less than the room temperature thermal energy of a lattice atom which disturbs the orderly spin arrangement. It follows then that the magnetic spin interaction is incapable of effecting their parallel orientation, characteristic of the ferromagnetics, at temperatures below the Curie point, and that the origin of the molecular field, which effects such parallel spin orientation, should be nonmagnetic. Subsequently, this conclusion was proved by direct experiments of Ya. Dorfman.

The role of exchange interaction in ferromagnetism. In 1928, Frenkel made the assumption that the origin of the forces which are responsible for the definite mutual orientation of atomic magnetic moments is electrostatic. They are the result of exchange interaction of the electrons of inner incomplete atomic shells. We have already discussed this type of interaction in dealing with the nature of the covalent bond (see Sec. § 3). The exchange interaction involves a change in the energy of the system. This is easily seen from the example of the simplest system of two hydrogen atoms (see Figure 1.5). According to Eqs. (1.11) and (1.12) the energy of such a system is

$$U = 2E_0 + \left(\frac{K \pm A}{1 \pm S^2} \right) \quad (7.70)$$

where E_0 is the energy of two noninteracting hydrogen atoms, K is the energy of Coulomb interaction of electric charges making up the atoms, S is the overlap integral whose value lies in the range $0 \ll S \ll 1$, and A is the exchange energy (in Chapter 5 we called it the exchange integral).

Calculation shows that A can be expressed by the following relation

$$A = -J(\mathbf{S}_i \cdot \mathbf{S}_j) \quad (7.71)$$

where \mathbf{S}_i and \mathbf{S}_j are the total spins of the interacting atoms, and J is the exchange integral (it is a measure of the probability of electron 1 going over to atom B and of electron 2 going over to atom A). In the case of two interacting hydrogen atoms

$$J = \int \left(\frac{q^2}{r} + \frac{q^2}{r_{12}} - \frac{q^2}{r_{b1}} - \frac{q^2}{r_{a2}} \right) \psi_a(1)\psi_b(2)\psi_a(2)\psi_b(1) dV_1 dV_2. \quad (7.72)$$

Here, q^2/r and q^2/r_{12} are the interaction energies of the nuclei between themselves and of the electrons between themselves, respectively; $-q^2/r_{b1}$ and $-q^2/r_{a2}$ are the energies of attraction of electron 1 to nucleus b and of electron 2 to nucleus a; $\psi_a(1)$ and $\psi_b(2)$ are the wave functions that describe the motion of electrons 1 and 2 around nuclei a and b, respectively; $\psi_a(2)$ and $\psi_b(1)$ are the wave functions that describe the probabilities for electrons 2 and 1 to be close to nuclei a and b, respectively, that is, the probabilities of the atoms A and B exchanging electrons; and dV_1 and dV_2 are volume elements.

It follows from Eq. (7.72) that both positive and negative terms enter the exchange integral. Therefore, the sign of the exchange integral may be either positive or negative. This is determined by the part played by the positive and negative terms of the exchange integral, which in its turn depends on the relation of the dimensions of the electron shells taking part in the formation of the exchange bond and on the interatomic distance.

The sign of the exchange integral determines what orientation of the spins of electrons taking part in the exchange bond is advantageous—the parallel or the antiparallel. It follows from Eq. (7.71) that when the sign of the exchange integral is negative ($J < 0$), the exchange energy A will be negative and, consequently, the system's energy U will be less than the energy $2E_0$ of the individual atoms [see Eq. (7.70)] if the spins \mathbf{S}_i and \mathbf{S}_j of the electrons taking part in the exchange bond are antiparallel: $\mathbf{S}_i \downarrow \uparrow \mathbf{S}_j$. As has been mentioned in Chapter 1, this case corresponds to the formation of a chemical bond between the atoms and the creation of a molecule [the symmetrical state described by Eq. (1.11)]; below we shall see that this is also a necessary condition for antiferromagnetism.

When the sign of the exchange integral is positive ($J > 0$), the exchange energy A will be negative and the energy of the system as a whole will be less than the energy of the individual atoms if the spins \mathbf{S}_i and \mathbf{S}_j of the electrons taking part in the exchange bond are parallel: $\mathbf{S}_i \downarrow \downarrow \mathbf{S}_j$. Hence, the parallel orientation of the spins of neighbouring atoms may too be advantageous, from the energy point of view, if their exchange integral is negative. This is the necessary condition for ferromagnetism since the parallel arrangement of spins and, consequently, of spin magnetic moments results in spontaneous magnetization, which is characteristic of ferromagnetics (see Figure 7.15).

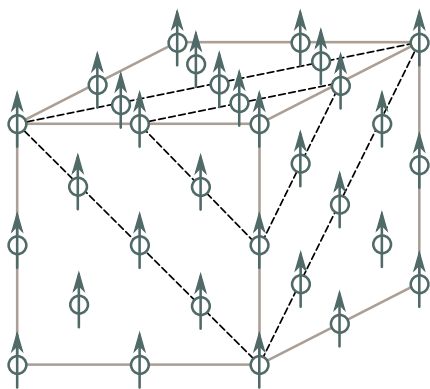


Figure 7.15: Spontaneous magnetization of a ferromagnetic. Exchange forces cause parallel orientation of the spins of electrons belonging to inner partially filled shells.

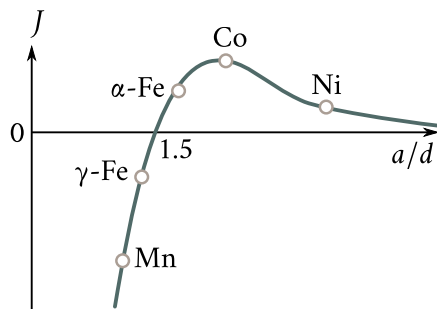


Figure 7.16: Dependence of exchange integral on the ratio of the lattice parameter to the diameter of the inner partially filled 3d shell in transition elements of the iron group.

Figure 7.16 shows the dependence of the exchange integral J on the ratio of the lattice constant a to the diameter d of the 3d shell of atoms of the iron group transition metals. It may be seen from Figure 7.16 that at $a/d > 1.5$, the exchange integral is positive; at $a/d < 1.5$ it turns negative, its absolute value increasing with the decrease in a/d . It follows then, that of all the transition metals only iron, cobalt, and nickel should be ferromagnetic, which is indeed the case. Manganese and other elements of the group, for which $a/d < 1.5$, are not ferromagnetics. Should it, however, be possible to increase somewhat the lattice constant of manganese so that the ratio a/d would approach 1.5, one could expect manganese to become a ferromagnetic.

Experiments support this view. For instance, inclusion of small amounts of nitrogen into the manganese lattice increases its lattice parameter and results in the appearance of ferromagnetism. Ferromagnetic properties are also exhibited by the Mn-Cu-Al alloys (Heusler alloys) and by the compounds like MnSb and MnBi in which the distances between the manganese atoms are greater than in pure manganese crystals.

Hence, the necessary and adequate conditions for ferromagnetism are the existence of incomplete internal atomic shells and the positive sign of the exchange integral which cause the parallel orientation of the spins.

Domain structure of ferromagnetic substances. Let us isolate a region A inside a ferromagnetic crystal [Figure 7.17(a)]. Suppose that exchange forces establish a parallel orientation of spins of all the electrons of incomplete atomic shells,

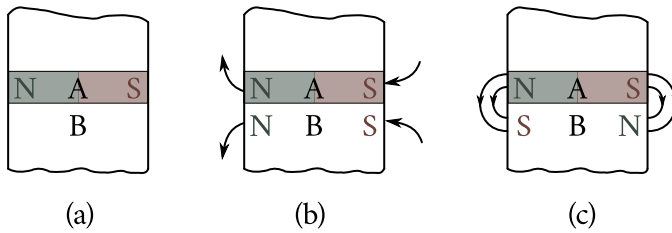


Figure 7.17: Ferromagnetic divided into domains (regions of spontaneous magnetization).

as shown in Figure 7.15. Region A will be magnetized to saturation. What will be the equilibrium spin orientation in the region B below region A? If the spins in region B are oriented as in A, there are two magnets with like poles in contact with each other [Figure 7.17(b)]. Such a state is unstable since it is characterized by the maximum energy of magnetic interaction. The stable state will be that in which the magnetic fields of the contacting regions are joined together, that is, a state in which the magnetization of the neighbouring regions of the crystal is opposite [Figure 7.17(c)]. Calculations show that as long as the width of region A does not exceed several interatomic distances, the dominant factor is the first—the orientation action of the exchange forces—whose effect is to magnetize the layers of region B in contact with region A in the same direction as that of region A. As the area of A widens, the importance of the second factor (the increase in the energy of magnetic interaction) grows and finally it becomes predominant: the width of region A reaches a critical value and the magnetization of the neighbouring region B from now on proceeds in the opposite direction. The critical width of the region of spontaneous magnetization is dependent on many factors, but usually it does not exceed several micrometers.

Thus, in the absence of an external field a ferromagnetic crystal should consist of a great number of separate and rather small regions magnetized to saturation. Those regions have received the name of *regions of spontaneous magnetization*, or *domains*. Domains are separated by layers in which the orientation of the spins changes from that of one domain to that of the other (Figure 7.18). Such transitional layers between domains became known as *Bloch walls*. In iron, their thickness is about 300 lattice constants (some 1000 Å). Figure 7.19 shows the domain pattern of a ferromagnetic predicted theoretically (a) and a photograph of the domain structure of an edge of a ferrosilicon crystal (b); arrows indicate the directions of spontaneous magnetization in the neighbouring domains.

Qualitative analysis of the magnetization curve. Spontaneous magnetization takes place in directions of easy magnetization. In the absence of an external field, the mutual orientation of the domains is such that the total magnetic moment

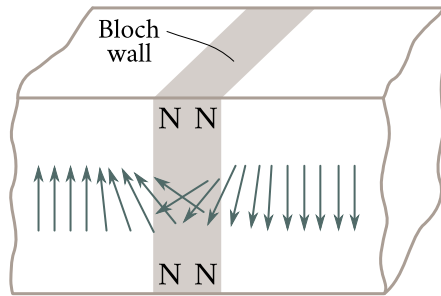


Figure 7.18: Structure of boundary layer separating two domains ("Bloch walls"); N denotes poles on the sample's surface.

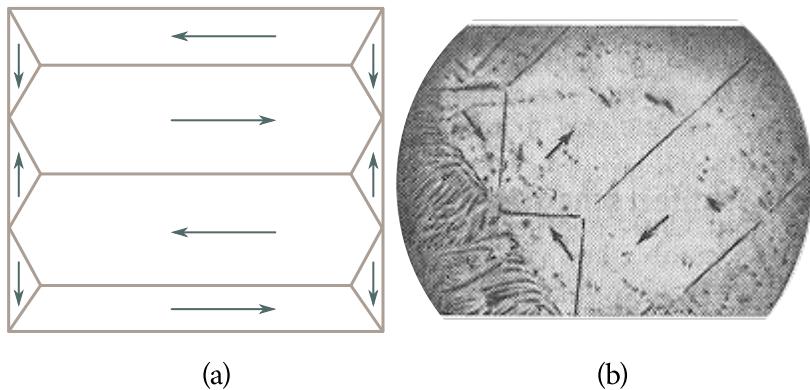


Figure 7.19: Domain structure of ferromagnetics: (a)—theoretically predicted pattern of a ferromagnetic's division into domains; (b)—photograph of an edge of ferrosilicon with decorated domain boundaries.

of the ferromagnetic as a whole is zero [Figure 7.20(a)], since this corresponds to the minimum of the system's free energy. When an external field \mathbf{H} is applied, the ferromagnetic is magnetized acquiring a nonzero magnetic moment. The nature of the physical phenomena which take place during the magnetization of a ferromagnetic is such that the process may be subdivided into three stages.

- (1) *Displacement of domain boundaries.* Place the crystal shown in Figure 7.20(a) in a magnetic field \mathbf{H} . The orientation of the magnetization vector \mathbf{J}_m of different domains with respect to \mathbf{H} is not the same: \mathbf{J}_m of the first domain makes the smallest angle with \mathbf{H} and that of the third domain the largest. When \mathbf{H} is increased it becomes advantageous from the viewpoint of energy for the most favourably oriented domain 1 to grow at the expense of domains 2, 3, and 4 [Figure 7.20(b)]. The mechanism of this growth is the displacement of the domain boundaries. For this reason, the first magneti-

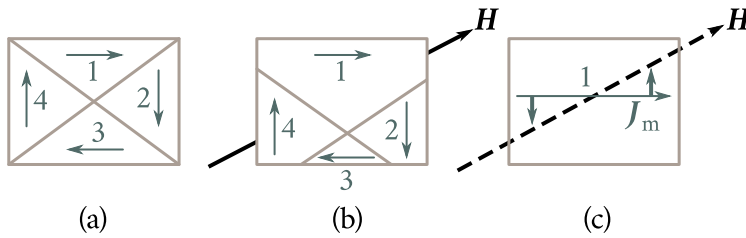


Figure 7.20: Processes involved in magnetization of a crystal: (a)—boundaries of four domains into which the crystal has been divided (the arrows indicate the direction of vector J_m); (b)—displacement of boundaries and growth of the most favourably oriented domain 1 with the increase in the magnetizing field H ; (c)—rotation of magnetization vector J_m in the direction of H .

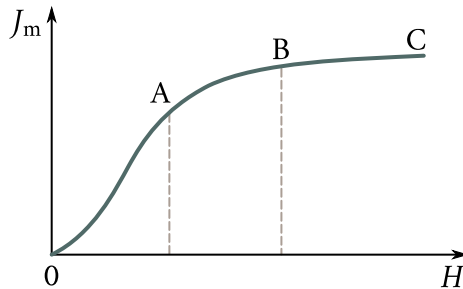


Figure 7.21: Magnetization plot of a ferromagnetic: OA—section corresponding to the process of displacement of domain boundaries, AB—section corresponding to rotation of magnetization vector, and BC—section corresponding to the paraprocess.

zation stage became known as the *displacement process*.

The displacement of the boundaries continues until the first domain spreads over the entire crystal. Figure 7.21 shows the magnetization curve of a single crystal. The displacement process is represented on this curve by the section OA. In small H 's, magnetization proceeds smoothly and is reversible; in strong fields it is a jumpy and an irreversible process leading to the Barkhausen effect.

- (2) *Rotation*. When H is further increased, the spontaneous magnetization J_m begins to rotate towards the field [Figure 7.20(c)]. The magnetization proceeds now at a much slower rate than in the first stage and ends when vector J_m coincides with H . At this stage the magnetization reaches technical saturation (Figure 7.21; section AB).
- (3) *Paraprocess*. After technical saturation is reached, the magnetization continues to grow with the increase in H although at a drastically reduced rate. The explanation is that, at any temperature other than absolute zero, not all the

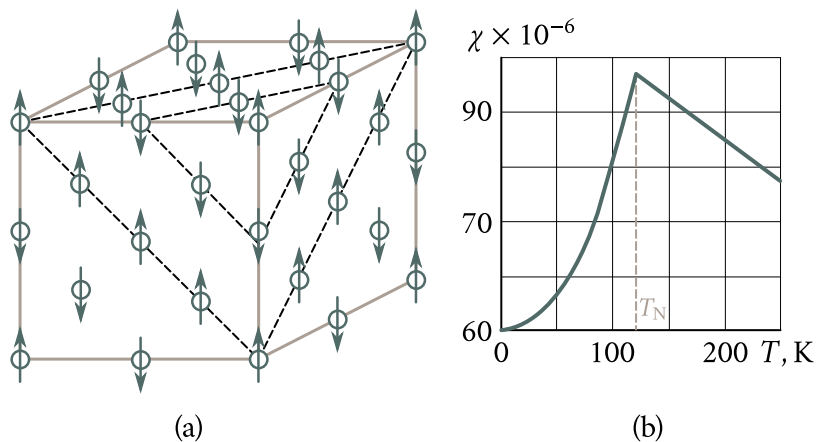


Figure 7.22: Magnetic structure of antiferromagnetic (a) and the temperature dependence of its magnetic susceptibility (b); the lattice of an antiferromagnetic (MnO) can be considered as consisting of two sublattices whose magnetic moments are antiparallel.

spins in the regions of spontaneous magnetization are oriented parallel to each other. Because of the thermal motion of atoms the orientation of some of the spins is antiparallel. The application of a strong magnetic field can effect the reorientation of these spins. The spin reorientation corresponding to the paraprocess is represented by the section BC.

§ 69. Antiferromagnetism

As was established in the preceding section, when the exchange integral is negative, the preferential orientation of the spins of neighbouring lattice sites is the antiparallel one. In this case, the spin arrangement can be also an ordered one, but there will be no spontaneous magnetization because the spin moments of the neighbouring lattice sites are antiparallel and compensate one another. Figure 7.22(a) shows the magnetic structure of MnO determined with the aid of neutron spectroscopy (only the magnetically active Mn atoms are shown in the figure). The structure may be regarded as a complex one consisting of two sublattices magnetized in opposite directions. Such structure can exist only below a certain temperature termed the *antiferromagnetic Curie point*, or the *Néel point*.

At absolute zero, the magnetic moments of the sublattices are mutually compensated and the total magnetic moment of the antiferromagnetic is zero. As the temperature is raised, the antiparallel arrangement of the spins is gradually disturbed and the magnetization of the antiferromagnetic rises; it reaches its maximum at the Néel point, at which the orderly spin arrangement vanishes altogether,

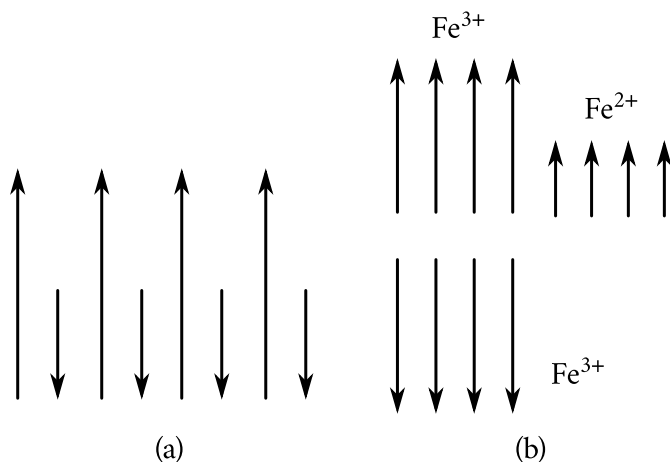


Figure 7.23: Schematic diagram of magnetic moments in a ferrimagnetic lattice in general (a) and specifically in magnetite $\text{FeO} \cdot \text{Fe}_2\text{O}_3$ (b). One of the sublattices is made up of the half of trivalent iron ions, the second sublattice being made up of the other half of trivalent iron ions and of bivalent ions of iron or of a substitute metal.

and the antiferromagnetic turns into a paramagnetic. As the temperature is raised still higher, the magnetization decreases in the same way as that of every paramagnetic. Figure 7.22(b) shows the temperature dependence of the magnetic susceptibility of MnO whose Néel point is $T_N \approx 120 \text{ K}$ in a field $H \approx 4 \times 10^4 \text{ A m}^{-1}$.

§ 70. Ferrimagnetism. Ferrites

The magnetic moments of the sublattices in antiferromagnetics are equal in magnitude and opposite in direction with the result that they completely compensate one another. However, there are cases when the magnitude of the magnetic moments of the sublattices is not the same owing, for instance, to the difference in the number or in the nature of atoms that make up the sublattices [Figure 7.23(a)]. This leads to the appearance of a finite difference in magnetic moments of the sublattices and to an appropriate spontaneous magnetization of the crystal. Such an uncompensated antiferromagnetism is termed *ferrimagnetism*.

The external behaviour of a ferrimagnetic is similar to that of a ferromagnetic, but because of the difference in their internal structure, the temperature dependence of their spontaneous magnetization may be quite different. For instance, the magnetization of a ferrimagnetic does not necessarily decrease monotonously with the rise in temperature but can pass through zero even before the Néel point is reached. Magnetite $\text{FeO} \cdot \text{Fe}_2\text{O}_3$ can serve as an example of a ferrimagnetic. The

negative oxygen ions form a face-centered cubic lattice in which there are one bivalent (Fe^{2+}) and two trivalent (Fe^{3+}) iron ions to every $\text{FeO} \cdot \text{Fe}_2\text{O}_3$ molecule. The bivalent iron ions may be replaced by bivalent ions of other metals, for instance, Mg, Ni, Co, Mn, Cu, etc., so that the general formula of materials of this class known as *ferrites*, assumes the form $\text{MeO} \cdot \text{Fe}_2\text{O}_3$, where Me stands for a bivalent metal. One of the sublattices of the complex ferrite lattice is made up of one half of the trivalent iron ions, and the other of the other half of trivalent iron ions and of bivalent ions of iron or the substitute metal. The magnetic moments of the sublattices are antiparallel. Therefore, the magnetic moments of the trivalent iron ions are mutually compensated and the magnetization is due to the magnetic moments of the bivalent metal ions [Figure 7.23(b)].

A remarkable property of ferrites is the combination of excellent magnetic parameters (high magnetic permeability, small coercive force, high saturation magnetization, etc.) with a high electrical resistance (of the order of $10^3 \Omega \text{ m}$). This particular property enabled ferrites to revolutionize the field of high and ultra-high frequency electronics. It is well known that ordinary low resistivity ($\approx 10^{-3} \Omega \text{ m}$) ferromagnetic materials cannot be used at such frequencies because of the extremely high eddy current losses. This was the reason why ferrites have occupied a unique position in this field.

Lately, ferrites with a high coercive force have been developed. They are used to construct permanent magnets capable of competing with electromagnets. Ferrites with a rectangular hysteresis loop are now widely used as digital storage elements in computers.

§ 71. Magnetic resonance

The magnetic moment of atoms, ions, and radicals with unpaired electrons is determined by relation (7.33). There are $2J + 1$ ways in which this moment may be aligned in a magnetic field H_0 , there being correspondingly $2J + 1$ different projections of the moment on the direction of the field. The energy corresponding to each such projection is

$$U_m = \mu_0 M_{JH} H_0 = m_J g \mu_0 \mu_B H_0. \quad (7.73)$$

Therefore, an atomic energy level splits in a magnetic field into $2J + 1$ sublevels (Figure 7.24) the separations between which are

$$\Delta U_m = g \mu_0 \mu_B H_0. \quad (7.74)$$

In the state of thermal equilibrium the atoms are distributed over those sub-

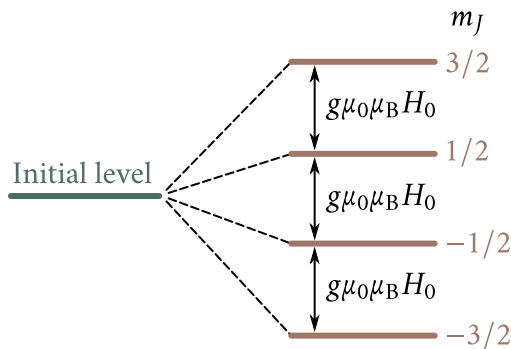


Figure 7.24: Splitting of a level of an atom with $J = 3/2$ in a magnetic field.

levels in accordance with the Boltzmann law:

$$n_1 = C \exp\left(-\frac{Jg\mu_0\mu_B H_0}{k_B T}\right), \quad n_2 = C \exp\left[-\frac{(J-1)g\mu_0\mu_B H_0}{k_B T}\right],$$

where n_1 is the number of atoms occupying the level with $m_J = J$, and n_2 the level with $m_J = J - 1$.

To effect transitions of the atoms from the lower to the higher sublevels an external electromagnetic field can be used. Spectroscopic selection rules allow only of such transitions which result in a unit change in the magnetic quantum number:

$$\Delta m_J = \pm 1 \quad (7.75)$$

that is, only transitions between adjacent sublevels the energy difference between which is $g\mu_0\mu_B H_0$. Such transitions can be excited by an electromagnetic field whose energy quanta are

$$\hbar\omega = g\mu_0\mu_B H_0. \quad (7.76)$$

Since transitions from the lower to the higher levels require a supply of energy, an intense absorption of electromagnetic energy will set in if condition (7.76) is fulfilled.

Condition (7.76) is the condition for *electron paramagnetic resonance* (EPR). The resonance frequency, as implied by (7.76), is a function of the constant magnetic field intensity H_0 . At $H_0 \approx 5.6 \times 10^5 \text{ A m}^{-1}$, $\nu_{\text{res}} \approx 2 \times 10^4 \text{ MHz}$, which corresponds to the wavelength $\lambda \approx 0.016 \text{ m}$.

A similar phenomenon is the *nuclear magnetic resonance* (NMR) due to the nuclear magnetic moment. For instance, in the case of protons, the nuclear resonance for $H_0 \approx 5.6 \times 10^4 \text{ A m}^{-1}$ occurs at a frequency of $\nu_{\text{res}} \approx 30 \text{ MHz}$, corresponding to a wavelength of electromagnetic radiation $\lambda \approx 10 \text{ m}$.

The first successful experiments on electron paramagnetic resonance were car-

ried out by E. K. Zavoisky in 1944. He measured the losses of electromagnetic energy in an electrical circuit caused by paramagnetic absorption. In 1945 H. C. Torrey and R. V. Pound used Zavoisky's method for the first successful experiments on the nuclear resonance of protons in solid paraffin. That moment marked the start of a rapid development of microwave spectroscopy—a formidable branch of physics dealing with the interaction of radiowaves with matter.

The magnetic resonance is extremely widely used in different fields of science and technology.

The nuclear magnetic resonance is the main and the most accurate method for measuring the magnetic moments of atomic nuclei. NMR has been helpful in collecting the data on the structure of liquids, dielectric crystals, metals, semiconductors, and polymers. The first investigations of population inversion of energy levels utilized in lasers were carried out with its aid.

The electron paramagnetic resonance makes it possible to study particles possessing unpaired electrons and processes in which such particles take part. Those particles include the conduction electrons, the free and bonded radicals, many atoms and ions. EPR is successfully applied in the study of the mechanisms of chemical reactions, the radiation effects in matter and in live tissues, the electronic state of solids (metals, dielectrics, and semiconductors), and in many other important fields of science and technology.

§ 72. Fundamentals of quantum electronics

Stimulated radiation. “Negative” absolute temperatures. The development of microwave spectroscopy in recent years, led to one of the most momentous technical discoveries—the discovery of the field *quantum electronics*—based on the ideas first put forward by Soviet physicists V. A. Fabrikant, N. G. Basov, and A. M. Prokhorov. Let us take a look at those ideas.

An external radiation directed at a quantum system causes not only the transitions from the lower to the upper levels, in which the energy is absorbed, but also the transitions from the higher to the lower levels, in which energy is liberated. Such radiation is termed *stimulated*.

Consider the simplest quantum system with only two levels (Figure 7.25). When N radiation quanta pass through a system, the difference Δ in the number of transitions, from the lower levels to the higher levels and from the higher levels to the lower levels, will be proportional to the transition probability w identical for both, the direct and the reverse processes, to the number of quanta N , and to the

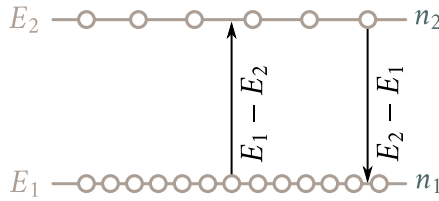


Figure 7.25: Two-level quantum system.

difference in the population of the levels ($n_2 - n_1$):

$$\Delta = wN(n_2 - n_1). \quad (7.77)$$

In conditions of thermal equilibrium, the distribution of the particles over the levels is described by the Boltzmann law:

$$n_1 = C e^{-E_1/k_B T}, \quad n_2 = C e^{-E_2/k_B T}, \quad , \quad \frac{n_2}{n_1} = \exp \left[-\frac{(E_2 - E_1)}{k_B T} \right]. \quad (7.78)$$

Since $E_2 > E_1$, it follows that $n_2 < n_1$, and because of that, resonance absorption exceeds stimulated radiation and the system absorbs incident electromagnetic energy eventually transforming it into heat.

For stimulated radiation to exceed resonance absorption, the thermal equilibrium of the system must be disrupted by raising the population of the higher levels above that of the lower levels, that is, to make $n_2 > n_1$. Such population is termed *population inversion*. Quantum states with population inversion may conveniently be described with the aid of the concept of negative absolute temperature. From Eq. (7.78) we have

$$T = -\frac{(E_2 - E_1)}{k_B \ln(n_2/n_1)}. \quad (7.79)$$

In equilibrium, $n_2 < n_1$, $E_2 - E_1 > 0$, and $T > 0$ K. In case of population inversion, $n_2 > n_1$, $E_2 - E_1 > 0$, and consequently $T < 0$ K. Figure 7.26 shows the occupancy of states at various temperatures. At $T = 0$ K, all the particles occupy the lowest level¹ and the system's energy is at its minimum ($E_{\min} = nE_1$). As the temperature rises, some of the particles go over to the higher level and the energy of the system increases. At $T = \infty$, the population of the levels is equilized and the system's energy reaches the maximum value it can have in equilibrium [$E_{\max} = n(E_1 + E_2)/2$]. At $T < 0$ K, the population of the higher level exceeds that of the lower level and, because of that, the energy of the systems turns out to be greater than E_{\max} . Hence, the energy region corresponding to negative temperatures lies not below the absolute zero, as would appear at the first glance, but

¹To avoid contradiction with the Pauli exclusion principle, E_1 and E_2 must be taken to mean narrow energy bands: from E_1 to $E_1 + dE$ and from E_2 to $E_2 + dE$.

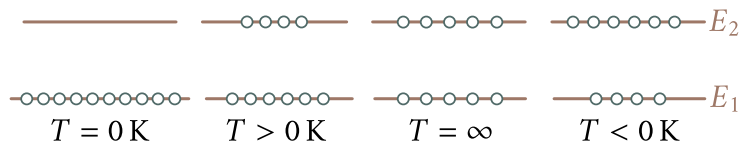


Figure 7.26: Level population at various temperatures.



Figure 7.27: Population inversion in a two-level quantum system.

above infinite temperature. Should such a system be brought in contact with a body whose temperature is positive, the heat would be transported from the system to the body until a state of thermal equilibrium would be reached.

It should be pointed out that negative temperature is a purely quantum effect and may be observed only in systems with a limited set of levels.

One may realize the population inversion in practice, for instance, by quickly reversing the direction of a constant magnetic field, so that the time of reversal would be less than the relaxation time. With the field H_0 in the initial direction the population of the lower level is greater than that of the upper one; when the field is reversed, quickly the initial population of the levels remains unchanged but with respect to the new field H_0 direction it will be populated inversely (Figure 7.27).

Principles of operation of masers. Suppose that an external signal with the resonant frequency $\omega = (E_2 - E_1)/\hbar$ is applied to a two-level system with a population inversion. This signal will induce the transitions of the particles: $E_2 \rightarrow E_1$ and $E_1 \rightarrow E_2$. Since in the case of population inversion the number of particles on the higher level exceeds that on the lower level, and since the transition probabilities w_{12} and w_{21} are equal, the stimulated radiation $E_2 \rightarrow E_1$ will exceed resonance absorption $E_1 \rightarrow E_2$, and the signal will be amplified. Therefore, such a system will do the job of an amplifier of electromagnetic radiation. The term for it is *paramagnetic amplifier* (usually *maser*).

In practice, not two-level but three-level quantum systems are used for paramagnetic amplifiers. The active materials in them are diamagnetically diluted crystals of paramagnetic salts; in wide use are ruby crystals (Al_2O_3) doped with chromium and germanium ions. Figure 7.28(a) shows a quantum system with three levels: E_1 , E_2 , E_3 . The dotted line represents the dependence of the number of particles n on the energy E in thermal equilibrium. Such a system is placed in a magnetic field H_0 and irradiated with high-frequency electromagnetic radiation with a frequency

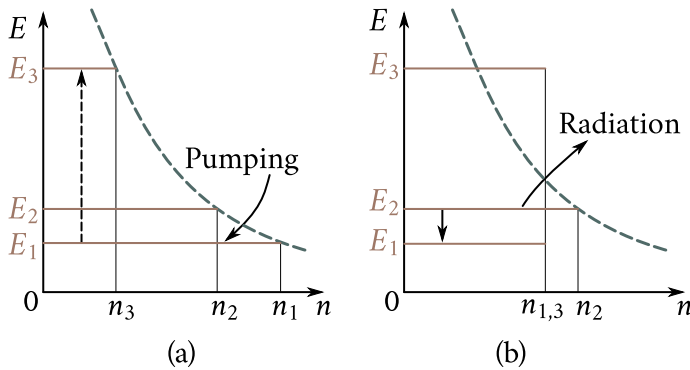


Figure 7.28: Excitation of paramagnetic amplifier (a) and radiation of amplified signal (b).

$\omega_{13} = (E_3 - E_1)/\hbar$. This process is termed *pumping*. In pumping fields of high intensity [Figure 7.28(b)], saturation is achieved, when the numbers of particles on the levels E_1 and E_3 are equal ($n_3 = n_1$) and less than that on the level E_2 : $n_1 < n_2$, $n_3 < n_2$. Should a signal with a frequency $\omega_{21} = (E_2 - E_1)/\hbar$ be applied to such a system, it would be amplified by the stimulated transitions of the particles from the level E_2 to the level E_1 . The system will work as an amplifier.

For a high amplification factor the difference in the population of the levels must be as great as possible. The way to do it is to cool the system to liquid helium temperatures. The concentration of the magnetic atoms should be low to exclude interaction of magnetic moments, which results in the widening of the absorption line and in the decrease in the amplification factor. One solution is to use diamagnetically diluted crystals of paramagnetic salts, of which the ruby crystal doped with chromium or germanium ions may serve as an example. The main advantage of the paramagnetic microwave frequency amplifiers is their ability to work at very low temperatures and, consequently, at low noise levels. This makes it possible to receive signals too weak for amplifiers of conventional types. The frequency tuning of the amplifier is done by changing the intensity of the field H_0 which in its turn changes the resonance absorption frequency.

Principles of operation of quantum generators. To devise a laser on the basis of the negative temperature quantum system, a positive feedback should be provided to make sustained oscillations possible. To this end, the system is placed inside a cavity with reflecting walls. In conditions in which each spontaneously generated quantum $\hbar\omega$ stimulates the generation of, on the average, more than one quantum, the amplitude of the electromagnetic oscillations of the appropriate frequency ω will grow continuously. The system will become self-excited. The radiative energy contained in the cavity is removed via a wave guide. Such cavities are

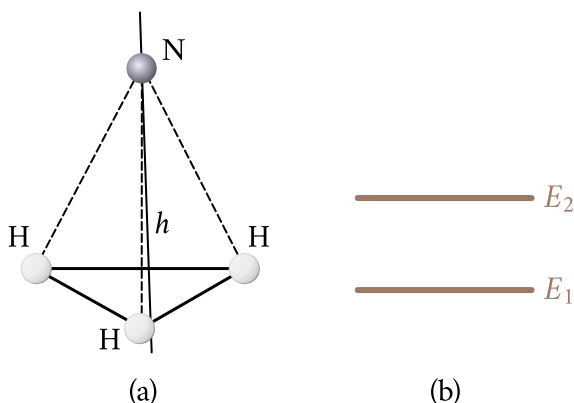


Figure 7.29: Ammonia molecule (a) and its energy levels (b).

termed *cavity resonators*. They are constructed from highly conductive materials and their dimensions are close to the wavelength radiated by the laser. The feedback signal is provided by radiation reflected by the cavity walls, a careful choice of the dimensions and shape of the resonator being necessary to obtain the right phase relationship (the reflected radiation must be in phase with the generated one at the point of generation).

The first quantum generators were designed by N. G. Basov and A. M. Prokhorov. They used two-level oscillators and the generator operated on beams of ammonia molecules. The ammonia molecule NH_3 is made up of three hydrogen atoms arranged in the base of a triangular pyramid and a nitrogen atom in the pyramid's vertex [Figure 7.29(a)]. This is the ground state of the molecule E_1 . The molecule has an excited state E_2 [Figure 7.29(b)] in which the nitrogen atom is forced into the base plane. The normal NH_3 molecule is asymmetric and by force of this has a nonzero dipole moment. The dipole moment of an excited molecule because of its symmetry is zero. This fact makes it possible to separate the normal and excited molecules by passing the molecular beam through a nonuniform electric field. Such a field deflects normal polar molecules; the excited nonpolar molecules are not deflected.

Figure 7.30(a) represents a schematic view of a molecular generator. It is made up of three parts: the beam source A, the separation system B, and the cavity resonator C.

The source of the molecular beam is a small space 1 closed on one side by a fine mesh 2. A gas pressure of one mmHg is maintained inside the space. The molecules forming the beam pass through the mesh into a vacuum chamber practically without collisions. A quadrupole condenser B [Figure 7.30(b)] is placed in

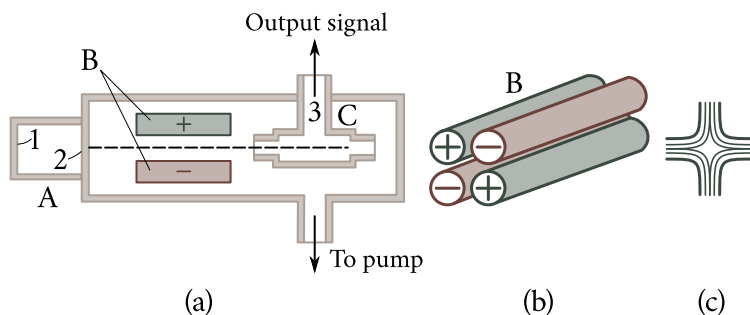


Figure 7.30: Schematic representation of a quantum generator operating on a beam of ammonia molecules: (a)—lay-out of amplifier (A is the source of beam of ammonia molecules, B the separation system, C the cavity resonator); (b)—quadrupole condenser of separating system; (c) electric field in condenser.

the way of the beam setting up a nonuniform field [Figure 7.30(c)] that sorts out the molecules. The excited molecules (on the higher level) are concentrated close to the condenser axis and the normal molecules are deflected to the walls. In this way, the beam close to the condenser axis is made to contain mainly the excited molecules and thus a quantum system with population inversion, or with negative temperature, is created.

From the separator B, the molecular beam enters the resonator C, turned to the frequency of the radiative transitions from the upper to the lower level [to the frequency $\omega = (E_2 - E_1)/\hbar$] and induces in it electromagnetic oscillations with this frequency. The electromagnetic energy is transported via the wave guide 3. The molecules in the beam practically do not interact and, because of that, the spectral line of the oscillations is very narrow (1 kHz at $\nu = 24$ MHz). Another advantage of the molecular generator is its long-term frequency stability. The oscillation frequency is determined solely by the structure of the ammonia molecule (that is, by $E_2 - E_1$) and is, therefore, independent of other generator circuit parameters. This made it possible to use the ammonia oscillator as a frequency standard. The error of a clock using such a frequency standard for its “pendulum” does not exceed one second in 300 years of continuous operation.

In recent years, quantum generators for the infrared and the optical spectral intervals (lasers) have been developed. Their use opens up wide possibilities for the development of communication, in locators, etc. For instance, an optical communication channel would be capable of carrying up to 10000 TV programmes. Quantum generators nowadays are widely used for the machining of tough materials, in medicine, and in other fields of science and technology.

Chapter 8

Contact Phenomena

§ 73. Work function

The work function concept. The positive ions that make up the metal lattice establish in it an electric field with a positive potential that changes periodically along a straight line passing through the lattice sites [Figure 8.1(a)]. As a rough approximation, this variation may be neglected and the potential be considered constant and equal to V_0 at every point of the metal. A free electron has a negative potential energy $U_0 = -qV_0$ in this field (q is the electron charge).

Figure 8.1(b) depicts the change in the potential energy of the electron as it passes from the vacuum into the metal: in vacuum $U = 0$, in the metal $U = U_0 = -qV_0$. Although such a change has the nature of a jump, it takes place over a distance δ equal approximately to the lattice parameter. It may be seen from Figure 8.1(b) that the metal is a potential trough for the electron and that work should be performed to get the electron out of it. This work is termed the *work function*.

Should the electrons have no kinetic energy, the work needed to liberate them would be equal to the depth of the potential trough U_0 . However, electrons possess kinetic energy of translational motion, even at absolute zero, since they occupy all the lower energy levels of the potential trough up to the Fermi level μ . Therefore, the energy needed to make them leave the metal is less than U_0 . The least work must be performed to liberate the electrons occupying levels close to the Fermi level. It is equal to the separation χ of the Fermi level from the zero level and the term for it is *thermodynamic work function*.

The problem of estimating the electron work function of a semiconductor is somewhat more complicated. As may be seen from Figure 8.2 the electrons may leave the semiconductor from the levels of the conduction band at the expense of work χ_0 , from the impurity levels at the expense of work χ_1 and from the levels of

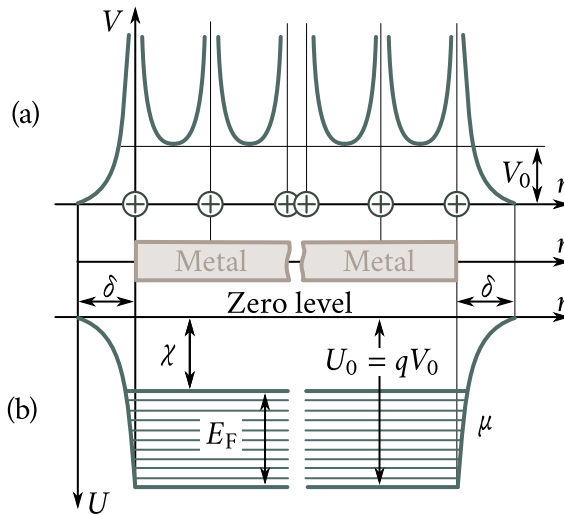


Figure 8.1: Metal as a potential through: (a)—internal potential of metal; (b)—potential energy of electron in metal.

the valence band at the expense of work χ_2 and χ_3 . The least work χ_0 is required to liberate the electrons from the conduction band. However, emission of only the conduction electrons would upset the equilibrium of the electron gas, to reestablish which the electrons should go over to the conduction band from the impurity levels and from the valence band. Such transitions require work to be performed and, in adiabatic conditions, this work is performed at the expense of the internal energy of the crystal, that is, as the state of thermal equilibrium is restored the crystal is cooled. If the electrons leave the semiconductor from the valence band, to restore equilibrium some electrons must go over from the conduction band to the valence band, which results in liberation of energy and the crystal being heated. The equilibrium will be maintained and the temperature will remain constant only if the electrons leave the semiconductor from the levels both below and above the Fermi level in appropriate numbers. Theory shows that to maintain equilibrium the average energy of the electrons leaving the semiconductor should be equal to the Fermi energy, and this is the work function, although there may be no electrons on the Fermi level itself.

The work function is measured in electron volts. The ratio of the work function to the electron charge is the voltage equivalent of the work function. The work function measured in electron volts is numerically equal to its voltage equivalent.

Effect of adsorbed layers on work function. Molecular layers adsorbed by the surface of the solid, in particular monomolecular layers, greatly affect the work

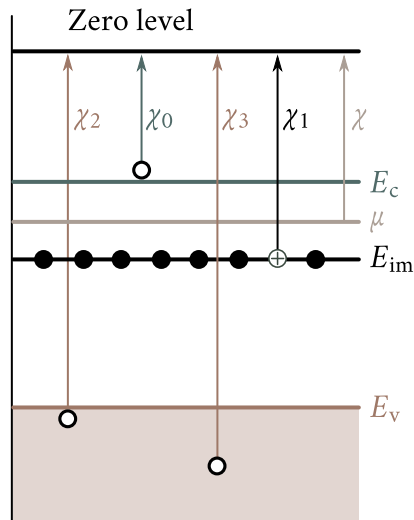


Figure 8.2: Electron work function of semiconductor.

function. Figure 8.3(a) shows a monatomic caesium layer on the surface of tungsten. Caesium is an alkali metal. Its outer valence electron is bonded to the nucleus much more weakly than the valence electrons of the tungsten atom. Therefore, in the process of adsorption caesium donates its valence electron to tungsten and turns into a positively charged ion inducing a negative change of equal magnitude in the metal's surface layer. When tungsten is covered by a monatomic caesium layer, an electric double layer is built up at the surface with its outer side charged positively. The potential difference in this double layer aids electron emission out of tungsten. Therefore, the electron work function, as determined from experiment, drops in the presence of the caesium layer from 4.52 eV for pure tungsten to 1.36 eV. The same is the effect of monatomic layers of other electropositive metals: barium, cerium, thorium, etc.

The reduction of the work function by the adsorption of electropositive metals is widely used for manufacturing vacuum tube cathodes, photocathodes, etc.

Quite different is the effect of oxygen adsorbed on the metal's surface. The valence electrons in the oxygen atom are bonded much more strongly than in metals. Therefore, in the process of adsorption the oxygen atom instead of donating electrons, accepts two electrons from the metal and turns it into a negatively charged ion. As a result, the outer side of the electric double layer becomes negatively charged [Figure 8.3(b)] and the resulting electric field prevents the electrons from leaving the metal, thereby, increasing the work function.

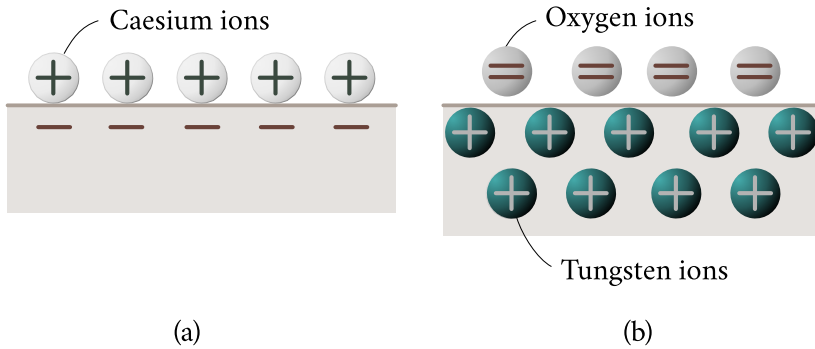


Figure 8.3: Formation of electric double layer in the course of adsorption of caesium (a) and of oxygen (b) on tungsten surface.

§ 74. Contact of two metals

Contact potential difference. Consider the process which takes place when two metals [Figure 8.4(a)] whose energy diagrams are shown in [Figure 8.4(b)] are brought together.

The electron gas in the individual metals 1 and 2 is characterized by the respective chemical potentials μ_1 and μ_2 , the thermodynamic work functions being χ_1 and χ_2 . Let us bring the metals closer together, to within such a distance d , that an effective electron exchange by means of thermionic emission or by direct transition from one metal to another, is possible. At the initial moment after the contact has been established, there will be no equilibrium between the electron gas in the first and second metals since the chemical potential (the Fermi level) μ_2 is above μ_1 . The difference in the Fermi levels, $\mu_2 - \mu_1$, results in the prevailing transition of the electrons from the second metal to the first one, in the course of which the first metal is charged negatively and the second positively. The appearance of the charges causes a shift in the energy levels of the metals: all the levels in the negatively charged metal 1 rise and in the positively charged metal 2 sink as compared with their positions in uncharged metals. This may easily be understood from the following simple considerations. To move an electron from, for instance, the zero level of an uncharged metal to the zero level of a metal charged negatively to a potential V_1 , work should be performed numerically equal to qV_1 . This work is transformed into the potential energy of the electron. Therefore, the potential energy of an electron occupying the zero level of the negatively charged metal will be higher, by the amount qV_1 , than the potential energy of an electron occupying the zero level of the uncharged metal. The zero level of a positively charged metal sinks below the zero level of an uncharged metal, the reason for this being the

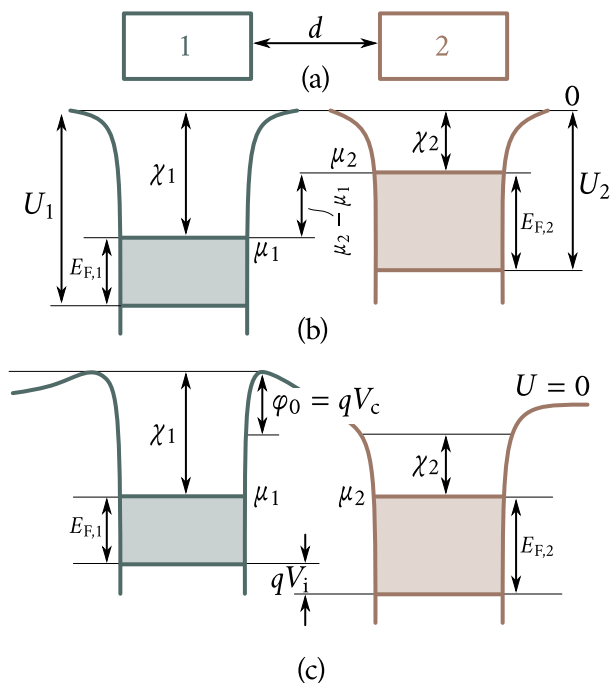


Figure 8.4: Origin of contact potential difference between n-type conductors.

same. A similar shift occurs in the position of other energy levels of the metals 1 and 2 including that of the Fermi level.

As soon as the continuously rising chemical potential of the metal 1 (μ_1) and the continuously sinking chemical potential of the metal 2 (μ_2) level out [Figure 8.4(c)], the cause for the predominant flow of the electrons from the first metal to the second disappears, and a dynamic equilibrium is established between the metals, resulting in the corresponding constant potential difference between the zero levels of both metals [Figure 8.4(c)] equal to

$$V_c = \frac{(\chi_1 - \chi_2)}{q}. \quad (8.1)$$

This potential difference is termed the *external contact potential difference*. It follows from Eq. (8.1) that it owes its existence to the difference in electron work functions of the contacting metals: the electrons leave the metal with the smaller work function and settle in the metal with the greater work function.

After the chemical potentials have been equalized, the kinetic energy of electrons occupying levels close to the Fermi levels of both metals is not the same: that of electrons in metal 1 is equal to E_{F1} and that of electrons in metal 2 to E_{F2}

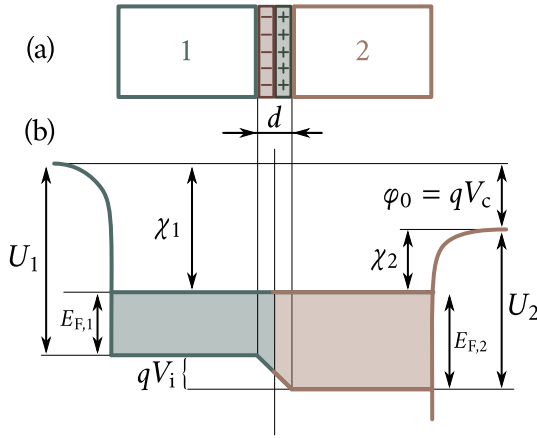


Figure 8.5: External (V_c) and internal (V_i) contact potential differences appearing at the instant two different metals are brought in contact.

($E_{F2} > E_{F1}$). When a direct contact is established between the two metals, a predominant diffusion process of the electrons from the second metal to the first sets in, and continues until the so-called *internal contact potential difference*

$$V_i = \frac{(E_{F2} - E_{F1})}{q}. \quad (8.2)$$

is established.

Thickness of the electric double layer at the contact of two metals. A double layer [Figure 8.5(a)] is established at the contact of two metals across which the potential changes abruptly by the amount V_1 [Figure 8.5(b)]. Let us estimate the thickness of this layer. Suppose that it is a plane condenser with the distance between the plates equal to the double layer thickness d . Denote the charge on each plate by Q and the potential difference by V_1 . The capacitance of a plane condenser with the plate area of 1 m^2 and a dielectric with relative permittivity $\varepsilon = 1$ is $C = \varepsilon_0/d$ (ε_0 is the permittivity of free space). Using the relation $C = Q/V_1$, we can rewrite this formula in the form $Q/V_1 = \varepsilon_0/d$. Hence, we obtain

$$d = \frac{\varepsilon_0 V_1}{Q}.$$

The thickness of the double layer cannot be less than the lattice parameter $a \approx 3 \text{ \AA}$. At $V_i \approx 1 \text{ eV}$ such a layer can be established if a charge $Q \approx V_i \varepsilon_0/a \approx 3 \times 10^{-2} \text{ C}$ is transported from every square metre of the contact surface of the first metal to the second. This corresponds to $\Delta n = Q/q \approx 2 \times 10^{17} \text{ m}^{-2}$ electrons. There are approximately 10^{19} atoms on every square metre of a metal. If we assume that each of them donates to the electron gas one valence electron, we will obtain

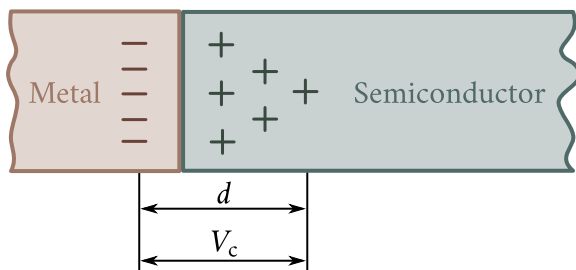


Figure 8.6: Formation of barrier layer in metal-semiconductor contact.

for the surface density of the electron gas the value $n_s \approx 10^{19} \text{ m}^{-2}$. Comparing Δn with n_s , we see that even for the narrowest possible double layer ($a \approx 3 \text{ \AA}$) only two percent of free electrons need be transported from the contact surface of one metal to another.

Because of a small change in the electron concentration in the contact layer, and because of the small thickness of this layer in comparison with the electron mean free path, the conductivity of the layer cannot be much less than that of the bulk metal. The current passes through the contact of two metals just as easily as through the metals themselves.

§ 75. The metal-semiconductor contact

Barrier layer. Consider a metal-semiconductor contact. Suppose that a metal with its work function equal to χ_m is brought in contact with an n-type semiconductor whose work function is χ_s (Figure 8.6).

If $\chi_m > \chi_s$ the electrons will flow out of the semiconductor into the metal until the chemical potentials μ_m and μ_s are equalized and a state of equilibrium is established. A contact potential difference V_c will be established between the metal and the semiconductor whose order of magnitude will be the same as that in the metal-metal contact (several volts). To establish such a contact potential difference approximately the same number of electrons should be transported from the semiconductor to the metal as in the case of the contact of two metals. For a lattice parameter of the semiconductor $a \approx 5 \text{ \AA}$ (germanium) and an electron gas concentration in it $n \approx 10^{21} \text{ m}^{-3}$, there will be $n_s \approx 10^{14}$ electrons on 1 m^2 of the semiconductor's surface. Therefore, the transport of $\Delta n \approx 10^{17}$ electrons entails the "depletion" of about 10^3 atomic layers of the semiconductor.

Hence, the equalization of the chemical potentials of a metal and a semiconductor in contact with one another necessarily involves the transport of electrons to the metal surface from a boundary layer of appreciable width d of the semicon-

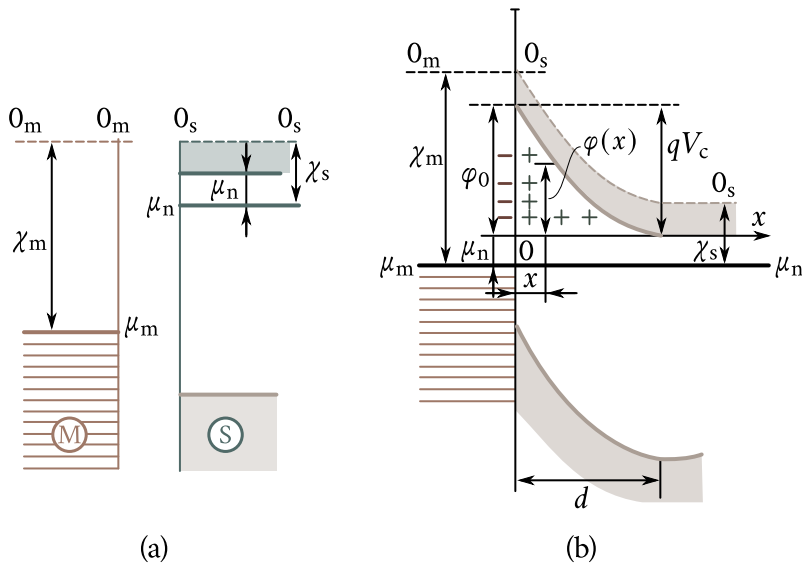


Figure 8.7: Effect of contact field on semiconductor's energy levels: (a)—band patterns of metal M and semiconductor S; (b)—deflection of semiconductor's energy bands by contact field.

ductor (Figure 8.6). The ionized impurity atoms remaining in this layer establish a static positive space charge. Since there are practically no free charge carriers inside this layer and since its width greatly exceeds the electron mean free path, its specific resistance is very great and because of that it is termed *barrier layer*.

Effect of contact field on semiconductor energy levels. The contact potential difference V_c between the metal and the semiconductor is built up over the entire width d of the barrier layer (Figure 8.6). Assuming that $a \approx 5 \text{ \AA}$, we obtain for the number of depleted atomic layers $\Delta N \approx 10^3$ and for the thickness of the barrier layer $d \approx 5 \times 10^3 \text{ \AA} \approx 10^{-7} \text{ m}$. For $V_c \approx 1 \text{ V}$, the field intensity will be $\mathcal{E}_c \approx V_c/d \approx 2 \times 10^6 \text{ V m}^{-1}$. This is at least three orders of magnitude less than the intensity of the internal crystal field, which is responsible for the energy band pattern of the semiconductor. Therefore, the contact field cannot appreciably affect the band spectrum (the forbidden band width, the impurity ionization energy, etc.). Its action is limited to the deflection of the semiconductor's energy bands. Let us dwell on this.

Figure 8.7(a) shows the energy band diagrams of the metal M and the semiconductor S before they have been brought in contact. When contact has been achieved and equilibrium has been established, a positive space charge throughout the barrier layer width d is built up [Figure 8.7(b)]. In the absence of the contact field, the

energy levels in the metal and in the semiconductor are represented by horizontal straight lines. This expresses the fact that the energy of an electron occupying this level, for instance, the lower level of the conduction band, is the same everywhere in the semiconductor and does not depend on the position of the electrons. In the presence of a contact potential difference the picture is changed: inside the layer in which the contact field is concentrated, a force acts on the electron pushing it out of the layer. To overcome this force, work should be performed, this work being equal to the potential energy of the electron in the contact field. Therefore, as the electron moves inside the space charge layer its potential energy $\varphi(x)$ increases reaching its maximum $\varphi_0 = qV_c$ at the semiconductor's surface. Quantum mechanical calculations lead to the conclusion that the application of an external field to the semiconductor results in an inclination of its energy bands in relation to the horizontal Fermi level. The contact field acts in the same way causing a deflection of the energy bands. The quantity φ_0 is termed the *equilibrium potential barrier* for electrons going over from the semiconductor to the metal.

To estimate the function $\varphi(x)$ we apply the Poisson equation known from electrostatics. This equation relates the field potential $V(x)$ to the density $\rho(x)$ of the static space charge responsible for this field. The equation is of the form

$$\frac{d^2V}{dx^2} = -\frac{\rho(x)}{\varepsilon_0\varepsilon} \quad (8.3)$$

where ε is the relative permittivity of the semiconductor.

It is expedient to go over from the potential $V(x)$ to the potential energy of the electron $\varphi(x) = -qV(x)$ and to write the Poisson equation for $\varphi(x)$ as

$$\frac{d^2\varphi}{dx^2} = \frac{q}{\varepsilon_0\varepsilon}\rho(x). \quad (8.4)$$

In calculating the space charge density $\rho(x)$ we shall assume all the donor atoms N_d to be ionized and their electrons to be transferred to the metal. Then $\rho(x) = qN_d$. Substituting this into the Poisson equation, we obtain

$$\frac{d^2\varphi}{dx^2} = \frac{q^2}{\varepsilon_0\varepsilon}N_d. \quad (8.5)$$

If we assume that there is no contact field at a distance $x \gg d$ inside the semiconductor, we will be able to write the boundary conditions for this equation in the form

$$\varphi(d) = 0, \quad \left[\frac{d\varphi(x)}{dx} \right]_{x=d} = 0. \quad (8.6)$$

Solution of Eq. (8.5) with the boundary conditions (8.6) yields:

$$\varphi(x) = \frac{q^2N_d}{2\varepsilon_0\varepsilon}(d-x)^2. \quad (8.7)$$

It follows from Eq. (8.7) that the potential of the contact field diminishes parabolically with the increase in x in the semiconductor.

For $x = 0$, we find that $\varphi_0 = \chi_m - \chi_s$. Substituting this into Eq. (8.7), we obtain the width of the barrier layer:

$$d = \left(\frac{2\varepsilon_0\varepsilon\varphi_0}{q^2N_d} \right)^{1/2} = \left(\frac{2\varepsilon_0\varepsilon V_c}{qn_{n0}} \right)^{1/2} \quad (8.8)$$

where $n_{n0} = N_d$ is the concentration of electrons (majority carriers) in the n-type semiconductor.

It follows from Eq. (8.8) that the thickness of the barrier layer d increases with the contact potential difference V_c determined by the difference in work functions and decreases with the concentration of majority carriers in the semiconductor. Table 8.1 shows the values of d calculated with the aid of Eq. (8.8) assuming that $V_c = 1$ V and $\varepsilon = 10$.

It follows from the table that for the electron gas concentrations typical of semiconductors (10^{20} m^{-3} to 10^{24} m^{-3}), the thickness of the surface layer containing practically no free carriers may attain values from one to three orders of magnitude greater than the electron mean free path. This is the reason why the resistance of the barrier layer is enormous.

If the work function of an n-type semiconductor exceeds that of a metal, $\chi_s > \chi_m$, the electrons shall be transported from the metal to the semiconductor setting up a negative charge in its contact layer (Figure 8.8). In this case, the energy of the electron $\varphi(x)$ as it approaches the surface does not increase but, on the contrary, decreases with the result that the bands are deflected in the opposite direction. This leads to an increase in the free charge carrier concentration inside the contact layer of the semiconductor and to a consequent increase in its conductivity. For this reason such a layer is termed *antibarrier*.

Rectification at a metal-semiconductor contact. A remarkable feature of the barrier layer is a drastic dependence of its resistance on the direction of external voltage applied to the contact. This dependence is so strong that it results

Table 8.1

$n_{n0} \text{ (m}^{-2}\text{)}$	$d \text{ (m)}$
10^{24}	3×10^{-7}
10^{22}	3×10^{-6}
10^{20}	3×10^{-5}

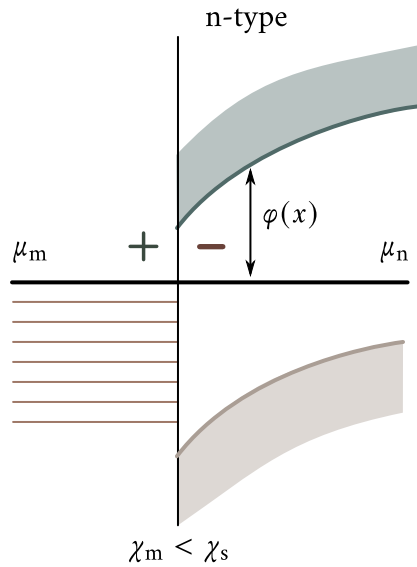


Figure 8.8: Formation of antibarrier layer in metal-semiconductor contact.

in practically a unidirectional conductivity of the contact: a current passes easily through the contact in the forward direction and much worse in the reverse direction. This is the essence of the rectifying property of a metal-semiconductor contact. Let us discuss this point in detail.

Figure 8.9(a) shows the energy band pattern of an n-type metalsemiconductor contact in the state of equilibrium. The potential barrier for the electrons going over from the metal to the semiconductor, is equal to the difference in work functions $\chi_m - \chi_s$; for the electrons passing from the semiconductor to the metal it is $\varphi_0 = qV_c$. Denote the electron flux from the metal to the semiconductor by $n_{s,m}^0$ and the flux from the semiconductor to the metal by $n_{m,s}^0$. The corresponding current densities flowing through the contact are $i_{s,m}$ and $i_{m,s}$. Since in the state of equilibrium the total current flowing through the contact is zero, it follows that $i_{s,m} = i_{m,s}$. Denote the current density corresponding to the equilibrium currents $i_{s,m}^0$ and $i_{m,s}^0$ by i_{eq} :

$$i_{s,m}^0 = i_{m,s}^0 = i_{eq}. \quad (8.9)$$

Apply an external potential difference V to the contact in the direction of the contact potential difference V_c imparting a positive charge to the semiconductor with respect to the metal [Figure 8.9(b)]; such direction of V is termed *reverse*. The resistance of the barrier layer is usually some orders of magnitude greater than the resistance of the other parts of the circuit and, because of that, practically the

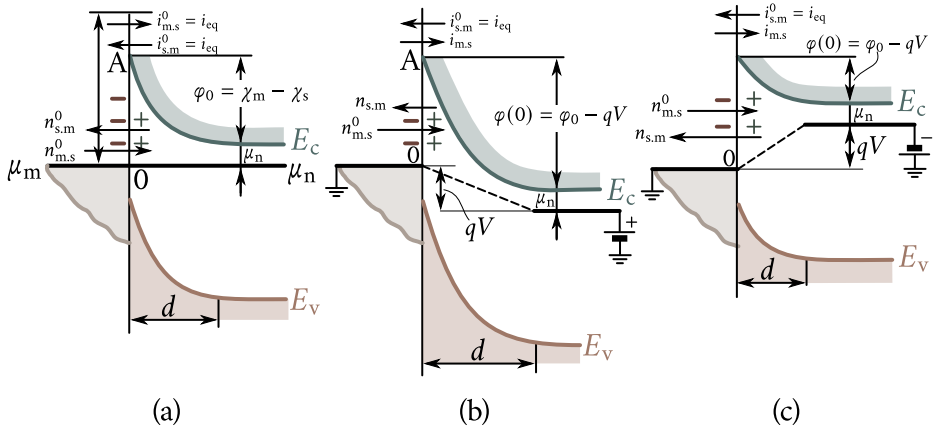


Figure 8.9: Rectification in metal-semiconductor contact: (a)—equilibrium state of contact; (b)—external voltage applied in reverse direction; (c)—external voltage applied in forward direction.

entire applied voltage is built up across the barrier layer. The energy levels in the positively charged semiconductor are deflected downwards by qV from their initial positions. The same will be the displacement of the Fermi level μ_n . The deflection takes place along the entire barrier layer thickness d across which the potential rises by V . The bottom of the conduction band E_c and the top of the valence band E_v in Figure 8.9(b) have been drawn so as to take account of the new position of the Fermi level. It may be seen from this drawing that the external voltage V applied in the reverse direction raises the potential barrier for the electrons going over from the semiconductor to the metal to

$$\phi(0) = \phi_0 + qV. \quad (8.10)$$

According to Eq. (8.8) the thickness of this barrier layer will be

$$d = \left[\frac{2\varepsilon_0\varepsilon(V_c + V)}{qn_{n0}} \right]^{1/2}. \quad (8.11)$$

Hence, the external field applied to the contact in the reverse direction raises the potential barrier for the electrons flowing from the semiconductor to the metal and increases the barrier layer width.

The picture will be different if a forward voltage is applied to the contact [Figure 8.9(c)]. In this case, all the levels of the negatively charged semiconductor including the Fermi level μ_n are deflected upwards by the amount qV , lowering the potential barrier for the electrons flowing from the semiconductor to the metal by the amount qV . The barrier height becomes

$$\phi(0) = \phi_0 - qV. \quad (8.12)$$

The width of the space charge layer decreases accordingly and becomes equal to

$$d = \left[\frac{2\epsilon_0\epsilon(V_c - V)}{qn_{n0}} \right]^{1/2}. \quad (8.13)$$

The change in the potential barrier height disturbs the equilibrium between the electron fluxes flowing through the contact in both directions. When the external voltage V is applied to the contact in the reverse direction, the current density $i_{m,s}$, corresponding to the electron flux $n_{s,m}$ decreases $e^{qV/(k_B T)}$ times, since according to the Boltzmann law the number of electrons flowing from the semiconductor to the metal capable of surmounting the barrier $\phi_0 + qV$ is $e^{qV/(k_B T)}$ times less than the number flowing through the equilibrium barrier ϕ_0 , that is, $n_{s,m} = n_{s,m}^0 e^{-qV/(k_B T)}$. Therefore, the current $i_{m,s}$ becomes equal to

$$i_{m,s} = i_{eq} e^{-qV/(k_B T)}.$$

The current density $i_{s,m}$, corresponding to the electron flux $n_{m,s}$, will remain equal to i_{eq} , since the external field does not change the height of the barrier for electrons flowing from the metal to the semiconductor: its height remains equal to the difference in work functions, $\chi_m - \chi_s$.

The total current density in the reverse direction is [Figure 8.9(b)]

$$i_r = i_{eq} e^{-qV/(k_B T)} - i_{eq} = i_{eq} \left[e^{-qV/(k_B T)} - 1 \right]. \quad (8.14)$$

The current flows from the semiconductor to the metal. As the reverse voltage V is increased, the exponential $e^{-qV/(k_B T)}$ tends rapidly to zero and the reverse current density to its limit value i_{eq} . The current density i_{eq} is termed *saturation current density* and $I_{eq} = i_{eq}S$ saturation current (S is the cross-sectional area of the metal-semiconductor contact).

When a forward external voltage V is applied [Figure 8.9(c)], the potential barrier for the electrons flowing from the semiconductor to the metal is lowered by the amount qV and, because of that, the current density of those electrons increases $e^{qV/(k_B T)}$ times in comparison with i_{eq} , becoming

$$i_{m,s} = i_{eq} e^{qV/(k_B T)}.$$

The current density $i_{s,m}$ remains equal to i_{eq} . Therefore, the density of the forward current (from the metal to the semiconductor) will be

$$i_r = i_{m,s} - i_{s,m} = i_{eq} \left[e^{qV/(k_B T)} - 1 \right] \quad (8.15)$$

and it grows exponentially with V . Combining Eq. (8.14) with (8.15), we obtain

$$i = i_{eq} \left[e^{\pm qV/(k_B T)} - 1 \right] \quad (8.16)$$

($V = |V|$ for forward bias and $V = -|V|$ for the reverse).

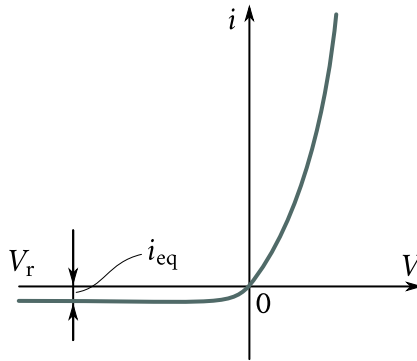


Figure 8.10: Current-voltage characteristic of metal-semiconductor contact.

Formula (8.16) is the equation of the current-voltage characteristic (CVC) of a rectifying metal-semiconductor contact. Figure 8.10 depicts the CVC of such a contact.

The ratio of the forward current to the reverse current for the same absolute value of applied voltage is termed *rectification ratio*. Its value for good rectifying contacts may be as high as tens and even hundreds of thousands.

The potential barrier at the metal-semiconductor interface is often termed *Schottky barrier*. Presently, Schottky barrier diodes with extremely short switching times are being developed. The diodes have switching times as short as 10^{-11} s. This makes it possible to use them effectively in radioelectronic pulse circuits, in computer and automation circuits where there is a need for high operational speeds, that is, where extremely short and quickly recurring electrical pulses have to be processed.

The nonrectifying (antibarrier) metal-semiconductor contact is used to provide ohmic contacts by means of which the semiconductor device is connected into the electrical circuit.

§ 76. Contact between two semiconductors of different types of conductivity

The progress in semiconductor electronics is based mainly on the use of contacts of two impurity semiconductors of different conductivity types. Such a contact is termed a *p-n junction*. Let us briefly discuss its properties.

Preparation of p-n junctions. It is impossible to obtain a true p-n junction by means of a mechanical contact of the n- and p-types of semiconductors because the lattice discontinuity at the interface contains more defects than there are impurity atoms on each of the contacting surfaces. Therefore, the p-n junction was

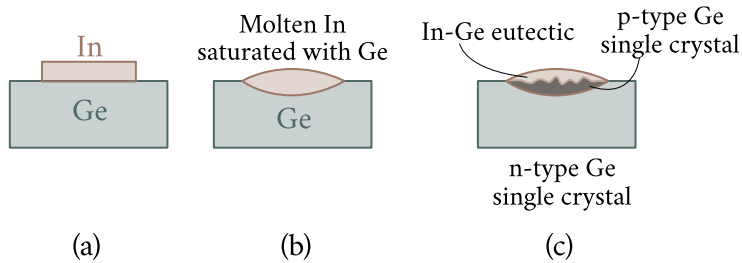


Figure 8.11: Fabrication of alloyed p-n junction: (a)—room temperature; (b)—temperature $T \approx 500^\circ\text{C}$; (c)—room temperature.

successfully prepared only when the art of making it in the form of an internal boundary in the bulk of a single crystal semiconductor was mastered.

One of the more widely used methods of preparing p-n junctions is the method of alloying, an example of which is discussed below. An n-type germanium wafer with a piece of indium placed on it [Figure 8.11(a)] is held in a furnace at a temperature $500\text{--}600^\circ\text{C}$ in a hydrogen or argon atmosphere. The indium melts and dissolves some germanium [Figure 8.11(b)]. As the wafer is slowly cooled, germanium saturated with indium precipitates from the melt. It crystallizes as a continuation of the single crystal of the wafer. Since the germanium doped with indium has a p-type conductivity, there will be a p-n junction at the boundary between the n-type single crystal that had not been dissolved and the recrystallized region [Figure 8.11(c)]. The indium drop alloyed to the germanium surface is used as an ohmic contact.

The p-n junction can be prepared by diffusing acceptor impurities into an n-type semiconductor or donor impurities into a p-type semiconductor. The diffusion process may be carried out from the gaseous, liquid, or solid phases. The penetration of the impurity and the depth of the p-n junction is determined by the temperature and the time of the diffusion process. The p-n junction itself is the boundary that separates the regions of different conductivity types.

A widely used method is the epitaxial method of preparing p-n junctions which consists in the deposition via chemical reactions on, for instance, an n-type silicon wafer of a p-type single crystalline silicon film in the gaseous phase or recrystallization from the liquid phase being used for the process.

A method growing in popularity in recent years is the ion implantation method, in which energetic ions of specific impurities (energy in the range of 50 keV to 300 keV) are directed at the semiconductor surface and penetrate into the bulk of it (to a depth of the order of $0.1\text{ }\mu\text{m}$ to $0.5\text{ }\mu\text{m}$, depending on the energy and the type of impurity).

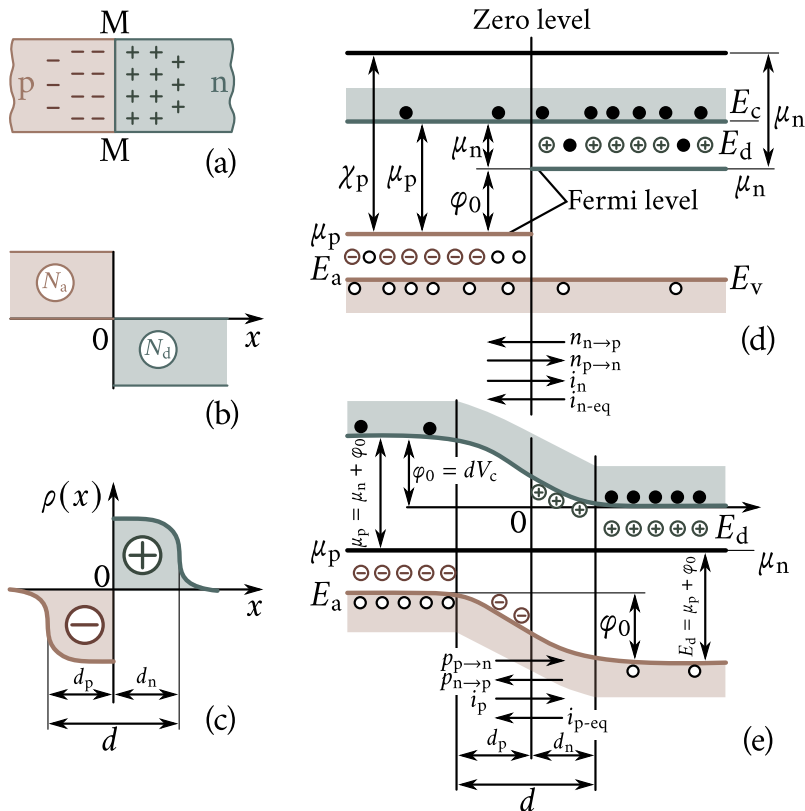


Figure 8.12: Equilibrium state of p-n junction: (a)—internal boundary MM between p- and n-regions; (b)—impurity distribution in p- and n-regions; (c)—distribution of immobile charges in p-n junction; (d)—position of Fermi levels at the time of imaginary contact of p- and n-regions; (e)—deflection of energy bands in the course of p-n junction formation and formation of a space charge layer.

Equilibrium state of a p-n junction. Let the plane MM be the internal boundary between two semiconductor regions of different conductivity type [Figure 8.12(a)]: to the left is the p-type semiconductor, for instance, p-germanium, with an acceptor concentration N_a , and to the right an n-type semiconductor (n-germanium) with a donor concentration N_d . For the sake of simplicity we shall assume $N_a = N_d$ to be equal to 10^{22} m^{-3} . Figure 8.12(b) depicts the change in the acceptor and donor concentrations along the x axis perpendicular to the plane MM. At point 0 lying in this plane, the acceptor concentration abruptly vanishes and the donor concentration increases abruptly from zero to N_d .

The majority carriers in the n-type region are electrons, and in the p-type re-

gion holes. The majority carriers are due almost entirely to ionized donor and acceptor impurity atoms. At temperatures outside the extreme low temperature range, practically all of those impurity atoms are ionized and, because of that, the electron concentration in the n-region (n_{n0}) can be assumed to be equal to the donor concentration N_d ($n_{n0} \approx N_d$), and the hole concentration in the p-region (p_{p0}) can be assumed to be equal to the concentration of the acceptor atoms ($p_{p0} \approx N_a$).

Besides majority carriers those regions contain also minority carriers: the n-region contains holes (p_{n0}) and the p-region electrons (n_{p0}). Their concentrations may be found from the law of mass action (5.44)

$$n_{n0}p_{n0} = p_{p0}n_{p0} = n_i^2$$

where n_i is the concentration of carriers of one sign in the intrinsic semiconductor (germanium). At $n_{n0} = p_{p0} = 10^{22} \text{ m}^{-3}$ and $n_i = 10^{19} \text{ m}^{-3}$ we get $p_{n0} = n_{p0} = 10^{16} \text{ m}^{-3}$.

We see that the hole concentration in the p-region is six orders of magnitude higher than in the n-region. Such a difference in the concentrations of carriers of one type is the cause of diffusion fluxes of electrons from the n-region to the p-region ($n_{n \rightarrow p}$) and of holes from the p-region to the re-region ($p_{p \rightarrow n}$). The diffusion flux of the electrons out of the n-region imparts a positive charge to this region, and the hole flux from the p-region imparts a negative charge to the p-region. Such charges raise the position of all the energy levels including the Fermi level in the p-region and sink it in the n-region. The electrons continue to flow from the right to the left and the holes from the left to the right until the gradually rising Fermi level of the p-region (μ_p) reaches the level of the gradually sinking Fermi level of the n-region (μ_n). As soon as those levels are equalized a state of equilibrium is established between the n- and p-regions when the electron flux from the n- to the p-region ($n_{n \rightarrow p}$) is compensated by the electron flux from the p- to the n-region ($n_{p \rightarrow n}$) and the hole flux from the p- to the n-region ($p_{p \rightarrow n}$) is compensated by the hole flux from the n- to the p-region ($p_{n \rightarrow p}$):

$$n_{n \rightarrow p} = n_{p \rightarrow n}, \quad p_{p \rightarrow n} = p_{n \rightarrow p}. \quad (8.17)$$

As the electrons leave the contact layer of the n-region, a static positive space charge of ionized donor atoms is left in it [Figure 8.12(c)]. Denote the width of this layer by d_n . As the holes leave the contact layer of the p-region, a static negative space charge of the ionized acceptor atoms is left there. Denote the width of this layer by d_p . A contact potential difference V_c is built up across those layers, which constitutes a potential barrier ϕ_0 localized in the p-n junction; ϕ_0 prevents the electrons from going over from the n- to the p-region and the holes from the p- to

the re-region. Calculations show that

$$\varphi_0 = k_B T \ln \left(\frac{n_{n0}}{n_{p0}} \right) = \ln \left(\frac{p_{p0}}{p_{n0}} \right). \quad (8.18)$$

It follows from Eq. (8.18) that φ_0 is the greater the greater the ratio of the majority carrier concentration in one region of the p-n junction to the concentration of the carriers of the same type in another region where they are minority carriers. At $n_{n0} = 10^{22} \text{ m}^{-3}$, $n_{p0} = 10^{16} \text{ m}^{-3}$, and $T = 300 \text{ K}$ we see that $\varphi_0 \approx 0.45 \text{ eV}$.

Figure 8.12(d) shows the energy band pattern of the p- and n-regions at the instant of their imaginary contact, that is, before equilibrium between them has been established. It may be seen from Figure 8.12(d) that μ_n lies above μ_p .

Figure 8.12(e) shows the energy band pattern of those regions after equilibrium has been established. The Fermi levels μ_n and μ_p coincide and there is a space charge layer between the p- and n-regions that spreads into the n-region to a depth d_n and into the p-region to a depth d_p forming a potential barrier with the height $\varphi_0 = qV_c$. Comparing Figures 8.12(d) and (e), one may easily see that

$$\varphi_0 = \mu_n - \mu_p. \quad (8.19)$$

The width of the space charge layer $d = d_n + d_p$, as in the case of the metal-semiconductor contact, is determined by the height of the potential barrier φ_0 and by the concentrations of majority carriers in both regions of the p-n junction n_{n0} and p_{p0} :

$$d = \left[\left(\frac{2\varepsilon\varepsilon_0\varphi_0}{q^2} \right) \left(\frac{n_{n0} + p_{p0}}{n_{n0}p_{p0}} \right) \right]^{1/2} = \left[\left(\frac{2\varepsilon\varepsilon_0V_c}{q^2} \right) \left(\frac{n_{n0} + p_{p0}}{n_{n0}p_{p0}} \right) \right]^{1/2}. \quad (8.20)$$

It may, however, be demonstrated that in the p-n junction the barrier width d depends eventually only on the majority carrier concentrations n_{n0} and p_{p0} . Indeed, substituting Eq. (5.38) into (8.18) and the result into Eq. (8.20), we obtain

$$d = \left[\left(\frac{2\varepsilon\varepsilon_0}{q^2} \right) \left(\frac{n_{n0} + p_{p0}}{n_{n0}p_{p0}} \right) k_B T \ln \left(\frac{n_{n0}p_{p0}}{n_i^2} \right) \right]^{1/2}.$$

It follows from Eq. (8.20) that the space charge layer width is the greater the less the majority carrier concentration in the n- and p-regions of the semiconductor.

If one of the regions, for instance the n-region, is substantially less doped than the p-region, so that $n_{n0} \ll p_{p0}$, we will obtain from Eq. (8.20) the following:

$$d \approx d_n \approx \left(\frac{2\varepsilon\varepsilon_0\varphi_0}{q^2 n_{n0}} \right)^{1/2}. \quad (8.21)$$

In this case, almost the entire space charge is concentrated in the low-doped (high resistivity) n-region, the same as in the case of a metal-semiconductor contact.

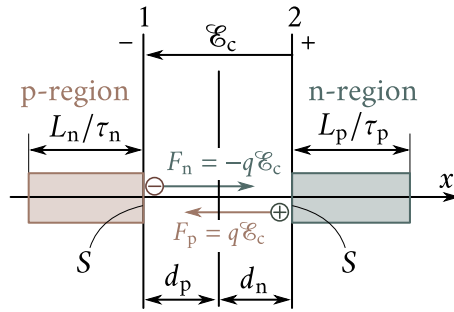


Figure 8.13: Calculating minority carrier currents flowing through equilibrium p-n junction.

Rectifying properties of p-n junctions. A remarkable property of the p-n junction, on which the operation of most semiconductor devices is based, is its ability to rectify alternating current. Let us discuss this property in more detail.

Currents flowing through p-n junction in equilibrium. In the state of equilibrium, fluxes of majority and minority carriers flow through the p-n junction [Figure 8.12(c)]. These fluxes are such that the flux of electrons—majority carriers—flowing from the n- to the p-region ($n_{n \rightarrow p}$) is, according to Eq. (8.18), equal to the flux of electrons—minority carriers—flowing from the p- to the n-region ($n_{p \rightarrow n}$). In the same way, the flux of holes—majority carriers—flowing from the p- to the n-region ($p_{p \rightarrow n}$) is equal to the flux of holes—minority carriers—flowing from the n- to the p-region ($p_{n \rightarrow p}$). Denote the current densities corresponding to those fluxes as follows: the flux $n_{n \rightarrow p}$ corresponds to i_n , $n_{p \rightarrow n}$ to i_p and $p_{n \rightarrow p}$ to $i_{p\text{-eq}}$. In accordance with Eq. (8.17) we can write

$$i_n = i_{n\text{-eq}}, \quad i_p = i_{p\text{-eq}}. \quad (8.22)$$

Adding the right- and left-hand sides of those equations, we obtain

$$i_n + i_p = i_{n\text{-eq}} + i_{p\text{-eq}}.$$

The left-hand side of the relation expresses the component of the full current due to the majority carriers, and the right hand side the component due to the minority carriers. The full current flowing through the p-n junction will evidently be zero:

$$i = (i_n + i_p) - (i_{n\text{-eq}} + i_{p\text{-eq}}) = 0. \quad (8.23)$$

Let us calculate $i_{n\text{-eq}}$ and $i_{p\text{-eq}}$. To this end, cut out a unit area S of the left boundary 1 of the p-n junction (Figure 8.13). Using it as a base, build a cylinder with the side L_n/τ_n , where L_n is the diffusion length of the electrons in the p-region and τ_n their average lifetime. Since the diffusion length is the average distance the carrier passes during its lifetime, the ratio L_n/τ_n , evidently, expresses the average speed of electrons diffusing from the bulk of the p-region, where their concen-

tration is n_{p0} to the boundary 1, where they are drawn into the contact field and transported to the n-region.

The number of electrons contained in the cylinder is equal to its volume L_n/τ_n multiplied by the electron concentration n_{p0} , that is $L_n n_{p0}/\tau_n$. All those electrons will pass through the unit area S in one second and will be transported to the n-region establishing a current with a density

$$i_{n\text{-eq}} = \frac{qL_n n_{p0}}{\tau_n}. \quad (8.24)$$

One may similarly calculate $i_{n\text{-eq}}$ by building a cylinder of unit base with the side equal to L_n/τ_n at the boundary 2 of the p-n junction:

$$i_{p\text{-eq}} = \frac{qL_p p_{n0}}{\tau_p}. \quad (8.25)$$

Hence, in the state of equilibrium in the p-n junction,

$$\begin{aligned} i_n &= i_{n\text{-eq}} = \frac{qL_n n_{p0}}{\tau_n} \\ i_p &= i_{p\text{-eq}} = \frac{qL_p p_{n0}}{\tau_p}. \end{aligned} \quad (8.26)$$

Forward current. Apply a forward voltage V to a p-n junction in a state of equilibrium [Figure 8.14(a)], by connecting the positive terminal of the power supply to the p-region and the negative terminal to the n-region [Figure 8.14(b)]. This voltage brings the potential barrier for the majority carriers down to $\varphi_0 - qV$. Therefore, the electron flux from the n- to the p-region ($n_{n \rightarrow p}$), and the hole flux from the p- to the n-region will increase $e^{qV/(k_B T)}$ times with the resulting similar increase in the majority carriers current densities i_n and i_p :

$$i_n = \frac{qL_n n_{p0}}{\tau_n} e^{qV/(k_B T)}, \quad i_p = \frac{qL_p p_{n0}}{\tau_p} e^{qV/(k_B T)}.$$

At the same time the current densities of minority carriers $i_{n\text{-eq}}$ and $i_{p\text{-eq}}$ whose magnitude is independent of the p-n junction's barrier height shall remain the same as expressed by formulae (8.26). Therefore, the total current flowing through the p-n junction, to which a forward voltage V has been applied and termed *forward current* i_f , will now be not zero but

$$i_f = (i_n + i_p) - (i_{n\text{-eq}} + i_{p\text{-eq}}) = q \left(\frac{L_n n_{p0}}{\tau_n} + \frac{L_p p_{n0}}{\tau_p} \right) \left[e^{qV/(k_B T)} - 1 \right]. \quad (8.27)$$

Reverse current. Apply now a reverse voltage $-V$ to the p-n junction, by connecting the negative terminal of the power supply to the p-region and the positive one to the n-region [Figure 8.14(c)]. This voltage raises the potential barrier of the p-n junction to $\varphi_0 + qV$, with the result that the fluxes of the majority carriers $n_{n \rightarrow p}$

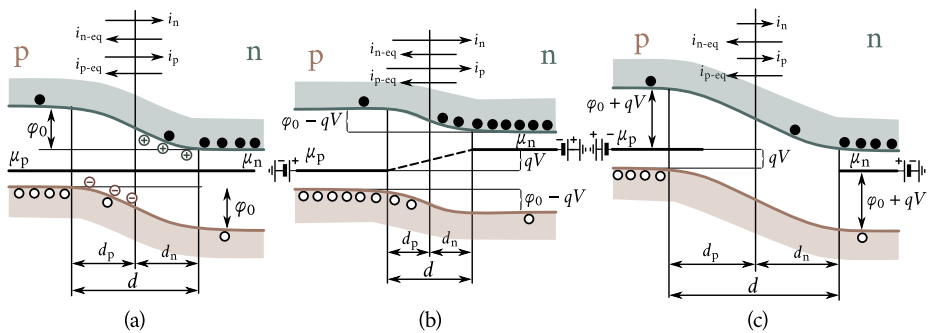


Figure 8.14: Rectifying action of p-n junction: (a)—equilibrium state of p-n junction; (b)—forward voltage applied to p-n junction; (c)—reverse voltage applied to p-n junction.

and $p_{p \rightarrow n}$ decrease $e^{qV/(k_B T)}$ times together with their currents i_n and i_p . The latter will be equal to:

$$i_n = \frac{qL_n n_{p0}}{\tau_n} e^{-qV/(k_B T)}, \quad i_p = i_{p-eq} = \frac{qL_p p_{n0}}{\tau_p} e^{-qV/(k_B T)}.$$

The total current flowing in the p-n junction termed *reverse current* i_r will be

$$i_r = (i_n + i_p) - (i_{n-eq} + i_{p-eq}) = q \left(\frac{L_n n_{p0}}{\tau_n} + \frac{L_p p_{n0}}{\tau_p} \right) \left[e^{-qV/(k_B T)} - 1 \right]. \quad (8.28)$$

Current-voltage characteristic (CVC). Combining Eqs. (8.27) and (8.28), we obtain the equation for the current-voltage characteristic of the p-n junction:

$$i = q \left(\frac{L_n n_{p0}}{\tau_n} + \frac{L_p p_{n0}}{\tau_p} \right) \left[\pm e^{-qV/(k_B T)} - 1 \right], \quad (8.29)$$

where $V > 0$ is the forward voltage, and $V < 0$ the reverse voltage.

Let us analyse this formula. With the increase in the reverse voltage $-V$, the exponent $e^{-qV/(k_B T)} \rightarrow 0$ and the expression $[e^{-qV/(k_B T)} - 1] \rightarrow -1$. Accordingly, the current density i_r tends to its limit

$$i_{eq} = -q \left(\frac{L_n n_{p0}}{\tau_n} + \frac{L_p p_{n0}}{\tau_p} \right), \quad (8.30)$$

termed *saturation current density*. Practically, this value is reached already at $qV \approx 4k_B T$, that is for $V \approx 0.1$ V. It follows from Eq. (8.30), that i_{eq} is determined by the minority carrier fluxes through the p-n junction. Since their concentrations are not large, i_{eq} is a small quantity. For germanium p-n junctions of the type discussed here ($n_{n0} \approx p_{p0} \approx 10^{22} \text{ m}^{-3}$) at room temperature, i_{eq} it is of the order of 10^{-2} A m^{-2} ; for silicon p-n junctions it is much less.

When a forward voltage V is applied to the p-n junction, the current density through it increases exponentially and reaches big values already at small voltages.

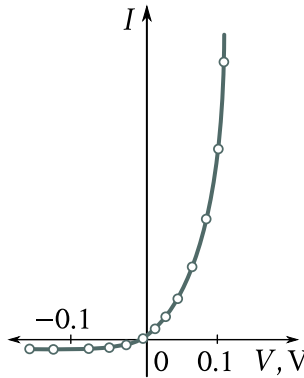


Figure 8.15: Current-voltage characteristic of p-n junction.

Substituting Eq. (8.30) into (8.29), we obtain:

$$i = i_{\text{eq}} \left[\pm e^{-qV/(k_B T)} - 1 \right]. \quad (8.31)$$

Figure 8.15 shows the plot of a current-voltage characteristic of a p-n junction which corresponds to equations Eqs. (8.29) and (8.31).

It is drawn to different scales for the forward and reverse branches since should the same scale be used for the reverse current as for the forward current the reverse branch would coincide with the x axis. Indeed, for $V_r = -0.5$ V, the reverse current density is $i_r \approx i_{\text{eq}}$, and for $V_f = 0.5$ V, the forward current density is $i_f \approx i_{\text{eq}} e^{20} \approx i_{\text{eq}} 10^9$, since at $T = 300$ K (room temperature) $k_B T \approx 0.025$ eV. As we see, the rectification coefficient at such a voltage is $i_f/i_r \approx 10^9$ and this proves that a p-n junction exhibits a practically unidirectional conductivity.

Deterioration of rectifying properties at high temperatures. According to Eq. (5.44)

$$p_{n0} = \frac{n_i^2}{n_{n0}}, \quad n_{p0} = \frac{n_i^2}{p_{p0}}$$

where

$$n_i = 2 \left(\frac{2\pi \sqrt{m_n m_p} k_B T}{h^2} \right)^{1/2} e^{-E_g/(k_B T)}.$$

It may easily be seen that n_i will rise rapidly with the increase in temperature while $n_{n0} \approx N_d$ and $p_{p0} \approx N_a$ will remain practically constant. Therefore, at some temperature n_i may become as high as n_{n0} or p_{p0} . Then, the concentrations of the minority carriers will be as high as the concentrations of the majority carriers: $p_{n0} = n_i^2/n_{n0} \approx n_{n0}^2/n_{n0} = n_{n0}$ and $n_{p0} = n_i^2/p_{p0} \approx p_{p0}^2/p_{p0} = p_{p0}$. The potential barrier in the p-n junction which is responsible for its rectifying properties will

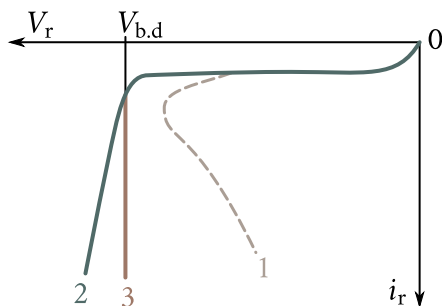


Figure 8.16: p-n junction breakdown: 1—thermal, 2—avalanche; 3—tunnel.

then cease to exist:

$$\varphi_0 = k_B T \ln \left(\frac{n_{n0}}{p_{p0}} \right) \approx k_B T \ln(1) \approx 0,$$

together with the ability of the junction to rectify alternating current. It follows from Eq. (5.37), that the corresponding temperature will be the higher the higher the forbidden band width E_g of the semiconductor. For germanium p-n junctions ($E_g = 0.62$ eV), the highest operational temperature is 75–90 °C; for silicon p-n junctions ($E_g = 1.12$ eV) it can be as high as 150 °C.

Breakdown in p-n junctions. If the reverse voltage is continuously increased, a voltage $V_{b,d}$ will be reached, at which the resistance of the barrier layer drops drastically and the reverse current jumps up. This phenomenon became known as the p-n junction breakdown (Figure 8.16).

There are different types of breakdown: thermal, tunnel (or Zener), and the avalanche breakdown in accordance with the nature of the physical processes that cause the reverse current to grow abruptly.

Thermal breakdown occurs when the heat generated by the reverse current flowing through the p-n junction is not completely removed from it and raises its temperature. A rising temperature leads to an increase in the reverse current and this, in its turn, raises the temperature still further, etc. This progressive process results eventually in thermal breakdown. The character of the increase in current during such a breakdown is depicted in Figure 8.16, curve 1.

When the electric field intensity in the p-n junction is high enough, impact ionization of the atoms of the semiconductor may take place. This will result in an avalanche-type increase in the carrier concentration and in the *avalanche breakdown*; the character of the increase in current is shown in Figure 8.16, curve 2.

In a narrow p-n junction already at comparatively low reverse voltages, an electric field may be established high enough for the tunnelling of the electrons through the p-n barrier to take place. The resulting breakdown is termed *tunnel*,

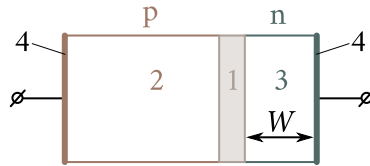


Figure 8.17: Schematic representation of a rectifier diode: 1—p-n junction, 2 and 3—passive regions, 4—ohmic contacts.

or *Zener, breakdown* and the character of current increase $i_r(V)$ for this case is shown in Figure 8.16, curve 3.

In most cases, the p-n junction breakdown is a harmful effect that limits the practical use of the junction. At the same time, the effect was utilized to develop a wide range of semiconductor devices known as Zener diode voltage regulators, which will be discussed in the following section.

§ 77. Physical principles of semiconductor p-n junction devices

As had already been stated before, the rapid progress in semiconductor electronics was the direct result of the development of the p-n junction technology on which the design of various semiconductor devices is based. Let us discuss the general principles of such devices.

Rectifier diodes. The nonlinear current-voltage characteristic of the p-n junction (Figure 8.15) enables it to be used to rectify alternating current. A two-terminal semiconductor device fulfilling such function is termed *semiconductor rectifier diode*. Figure 8.17 shows the principal schematic diagram of such a diode. It consists of a p-n junction 1, passive n- and p-regions 2 and 3, and ohmic contacts 4. The high-resistivity region of the crystal is termed *base* of the diode. In our case it is the n-region 3 of the width W .

Nowadays, p-n junction rectifier diodes are made almost exclusively of silicon. The efficiency of such diodes is almost 100 percent and in combination with their low weight, small dimensions, ease of servicing, etc. this made them a very widely used device for such applications. Various diode types are designed to rectify currents from several milliamperes to several hundred or thousand amperes. For greater currents the diodes are connected in parallel. The maximum reverse voltage for various types lies in the range from 50 V to 600 V. It may be much higher for special diode types. For use in high-voltage rectifiers the diodes are assembled in-series in stacks. The reverse currents of various rectifier types lie in the range from fractions of a microampere to tens of milliamperes.

Impulse and high-frequency diodes. The second very important field of ap-

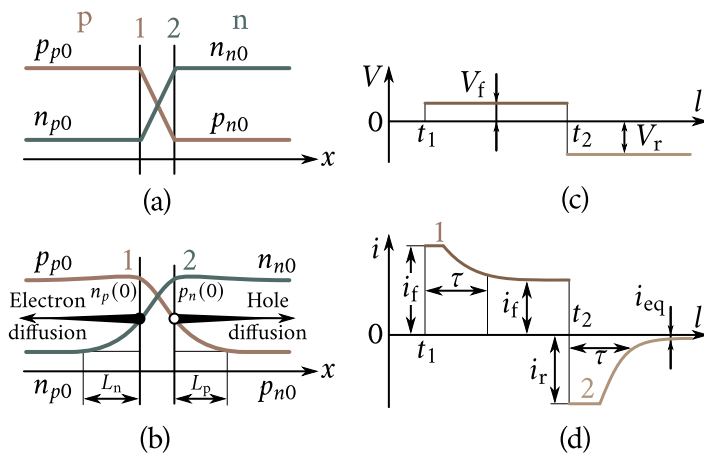


Figure 8.18: Transient processes in diode: (a)—majority and minority carrier distributions; (b)—minority carrier injection by forward voltage and their diffusion into the bulk of semiconductor; (c, d)—variations of forward and reverse voltages and currents in a diode switched on in the forward direction and reswitched to reverse direction.

plication of the semiconductor diodes is the field of impulse electronics, computer electronics, automation, VHF electronics, etc. In such applications, the diode is required to process pulses of minimum duration and of maximum repetition rate. Therefore, one of the main requirements of diodes designed for such circuits is the speed of operation, that is, switching speed from the direct to the reverse state. To find what lies at the origin of this speed let us discuss the physical processes that take place when a p-n junction is switched.

Figure 8.18(a) shows the distribution of the majority and minority carriers in the p- and n-regions of an equilibrium p-n junction. When a direct voltage V is applied to the diode, the potential barrier of its p-n junction sinks by qV and the majority carrier flux through the junction increases $e^{qV/k_B T}$ times, with the result that the hole concentration at boundary 1 and the electron concentration at boundary 2 rise to the values of $p_n(0) \gg p_{n0}$ and $n_p(0) \gg n_{p0}$, respectively [Figure 8.18(b)]. For the direct current to flow, it is necessary for those carriers to be drawn into the bulk of the semiconductor: the holes should be drawn into the n-region and the electrons into the p-region. The recombination of those minority carriers takes place inside the regions, or on the contacts if the width of the regions is small in comparison with their diffusion length. The removal of the minority carriers proceeds either by means of diffusion, whose rate is the higher the greater the concentration gradient of holes (dp/dx) at boundary 2 and that of electrons (dn/dx) at boundary 1 or in special diode types by means of drift in built-in electric fields. At the initial

moment after the forward voltage had been applied [Figure 8.18(c)], this gradient is extremely high, since the holes transported to the n-region and the electrons transported to the p-region are concentrated in narrow layers close to boundaries 2 and 1. Therefore, the initial forward current in the diode is high, being limited practically only by the resistance of its passive regions [plateau 1 in Figure 8.18(d)]. As the holes enter the n-region and the electrons the p-region, their concentration gradient drops and so does the forward current [Figure 8.18(e)]. After a time equal to the minority carrier lifetime τ (or to the transit time of the minority carriers from the boundaries 1 and 2 to the contacts 4, which is even shorter), a steady-state (independent in time) distribution of the holes in the n-region and of the electrons in the p-region is established [Figure 8.18(b)] and the forward current assumes its normal value [Figure 8.18(d)].

When the diode is switched from the forward to the reverse state [Figure 8.18(c)], the initial reverse current is very high since the minority carrier concentrations at boundaries 2 and 1, which are responsible for this current, are high: the magnitude of the current is actually limited by the resistance of the passive regions of the diode [plateau 2 in Figure 8.18(d)]. In the course of time, the excess carriers at the boundaries gradually dissolve, their concentrations drop to equilibrium values (p_{n0} and n_{p0}), and the reverse current assumes its normal value [Figure 8.18(d)]. This process lasts about the same time as the first (lifetime or transit time).

Thus, when the diode is switched, transient processes (of carrier accumulation in the forward biased diode and of carrier dissipation in the reverse biased diode) take place in it, limiting its switching speed. Since the duration of those processes is approximately equal to the minority carrier lifetime τ , and since a reduction in τ increases the switching speed, the tendency is to make τ as short as possible. Another method is to reduce the transit time of carriers by making the diodes as thin as possible (several microns or less).

On the basis of the aforesaid, we can draw the conclusion that, a p-n junction behaves with respect to a short alternating signal as a resistance R located in the barrier layer shunted by the capacitance C of the p-n junction (Figure 8.19). When a forward bias is applied, the initial current through the diode is mainly the current which charges the capacitor C . This current can be quite high. When the diode is switched to the reverse bias, the initial reverse current is mainly the discharge current of the capacitor C and it too can be quite high. To improve the speed of the diode and its high-frequency performance it is evidently necessary to reduce the p-n junction capacitance C . This is done by reducing the area of the p-n junction to the minimum. This measure together with other measures enabled the switching speeds of modern diodes to be reduced to approximately 10^{-9} s and the operating frequencies to be raised to 10^9 Hz.

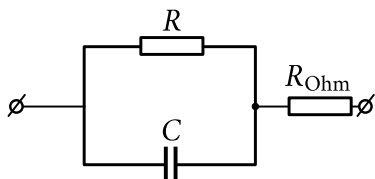


Figure 8.19: Equivalent circuit of diode: R —non-linear resistance of p-n junction, C —capacitance of p-n junction; R_{ohm} —ohmic resistance of passive regions and contacts.

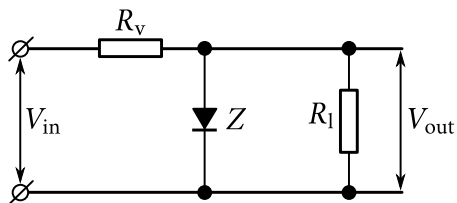


Figure 8.20: Circuit diagram of the simplest voltage regulator using a silicon Zener diode Z .

Voltage regulators. A small increase in the reverse voltage in the prebreakdown range causes a substantial increase in the reverse current (see Figure 8.16). This effect is being used to stabilize voltage and the device is called the *Zener diode regulator*.

Figure 8.20 depicts the simplest circuit diagram of a dc voltage stabilizer utilizing a Zener diode. When the input voltage F_{in} is increased, the diode's resistance drops drastically, and the current in the circuit of the voltage drop resistor R_v increases, causing the voltage drop across it to increase; the voltage across the load resistance R_l (the output voltage V_{out}) may remain practically constant. Power supplies using Zener diodes are now quite a match for normal cells.

Tunnel diodes. There is another very interesting and practically important type of semiconductor devices, the so-called *tunnel diode*, which utilizes the quantum mechanical effect of electrons tunnelling through a narrow potential barrier. The diodes are constructed from heavily doped degenerate semiconductor material in which the Fermi level lies not in the forbidden band, but just like in metals in the conduction band of an n-type semiconductor or in the valence band of a p-type semiconductor. Figure 8.21(a) shows the energy-band diagram of a tunnel diode in the state of equilibrium. We see that there is a partial overlapping of the valence band of the p-region and the conduction band of the n-region. This makes possible the tunnelling of electrons from the n-region to the p-region (flux 1) and from the p-region to the n-region (flux 2). Flux 1 constitutes the reverse tunnelling current and flux 2 the direct current. In the absence of an external field, those currents are equal and the total current through the junction is zero.

When a direct voltage is applied to the junction, the overlapping of the bands becomes smaller [Figure 8.21(b)] and, because of this, flux 2 exceeds flux 1 and a direct current passes through the junction increasing with the direct voltage V until a maximum is reached, corresponding to a voltage at which the bottom of the conduction band of the n-region coincides with the Fermi level of the p-region [Figure

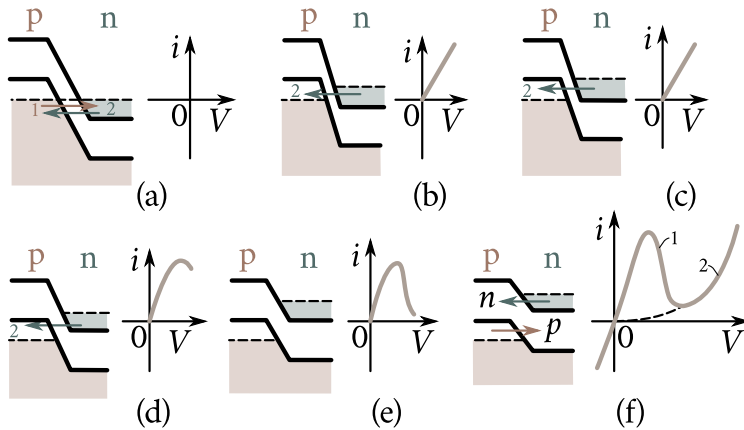


Figure 8.21: Principle of operation and current-voltage characteristic of the tunnel diode.

8.21(c)]. When V is increased still further, the direct current diminishes because of the decrease in the number of occupied states of the n-region lying opposite the free states of the p-region [Figure 8.21(d)]. When at a voltage V , the bottom of the conduction band of the n-region, coincides with the top of the valence band of the p-region, the overlapping of the bands ceases [Figure 8.21(e)] and the tunnel current turns zero, but a small direct current appears as in a usual diode. It rises rapidly with a further increase in V in accordance with Eq. (8.27) [Figure 8.21(f)].

A remarkable property of the tunnel diodes is the negative differential resistance region 1-2 on the current-voltage characteristic similar to that of the Gunn diode. This makes it possible to use those diodes for the generation of VHF oscillations up to frequencies of about 10^{11} Hz. The tunnel diode was one of the first devices with a switching time of a fraction of a nanosecond (10^{-10} s) and this enabled it to be used in impulse circuits of digital computers and in various automation circuits. Only the majority carriers work in the tunnel diodes and this makes them much less sensitive to ionizing radiation than the bipolar semiconductor devices, this fact being of special importance for space explorations.

The development of the tunnel diodes is an excellent illustration of the fact that the quantum mechanics, formerly an exotic science, became for the modern engineer a powerful tool, mastered in order to be able to take an active part in the progress of modern technology.

Transistors. Rapid progress in semiconductor electronics became possible only after the invention in 1948-1949 by J. Bardeen, W. H. Brattain, and W. Shockley of the semiconductor amplifier—the transistor—whose characteristics and whose designation were similar to those of the vacuum tube but which had some substantial advantages over the latter.

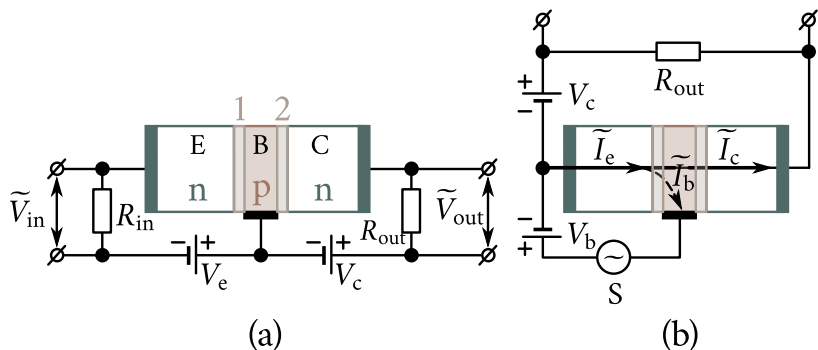


Figure 8.22: Schematic representation of n-p-n junction transistor and its connection into circuit.

Figure 8.22(a) shows the schematic representation of an n-p-n junction transistor. The transistor is made of three regions: the left n-region E termed *emitter*, the middle p-region B termed *base*, and the right n-region C termed *collector*. Those regions are separated by two p-n junctions: the emitter and collector junctions. By means of ohmic contacts the transistor is connected into the circuit: one of the possible connections (the common base connection) is shown in Figure 8.22(a), where R_{in} is the equivalent input resistance of the transistor and R_{out} its equivalent output resistance. It may be seen that the emitter p-n junction, is biased in the forward and the collector in the reverse direction.

To understand the physical principles of transistor operation let us turn again to Figure 8.18(b). When a forward bias is applied to the emitter p-n junction, the concentrations of the majority carriers—of electrons in the p-region and of holes in the n-region—increase drastically as compared with their equilibrium concentrations. This phenomenon is termed *minority carrier injection* and serves as the basis for transistor operation.

The electrons injected from the emitter E into the base B [Figure 8.22(a)] diffuse to the collector C; with a base that is narrow in comparison with the minority carrier diffusion length, practically all the injected electrons will reach the collector junction and will be drawn in by its field into the collector circuit of the transistor. Therefore, the collector current I_c will be approximately equal to the emitter current I_e : $I_c = \alpha I_e$, where $\alpha \approx 1$ is the common base current amplification factor.

Now imagine that an external signal V_{in} small in comparison with the bias voltage V is applied to the input resistance R_{in} . The input current—the emitter signal current—will be $I_e = V_{in}/R_{in}$ and the output voltage—the signal voltage across the collector junction (or the equivalent collector resistance R_{out})—will be $\tilde{V}_{out} = \tilde{I}_c R_{out} = \alpha \tilde{I}_e R_{out}$. Therefore, the voltage amplification factor α_V of the

transistor in the common base connection will be

$$\alpha_V = \frac{\tilde{V}_{in}}{\tilde{V}_{out}} = \alpha \frac{R_{out}}{R_{in}} \approx \frac{R_{out}}{R_{in}}.$$

Since R_{in} is a small differential resistance of a forward-biased p-n junction and R_{out} is an enormous resistance of a reverse-biased junction ($R_{out} > R_{in}$), $\alpha_V \gg 1$ and may be as high as 10^5 (for dc current). Since in this connection only the voltage is amplified, the same will be the power amplification factor $\alpha_P = P_{out}/P_{in} \approx \alpha_V$. The source of the additional signal power dissipated in the collector circuit is the collector power supply V_C .

Figure 8.22(b) shows the transistor connected into a common emitter circuit. In this case the signal from the source S is applied between the emitter and the base, the output signal being taken off the emitter and the collector. The input signal affects the emitter \tilde{I}_e , the collector \tilde{I}_c , and the base \tilde{I}_b currents, the latter being the difference of the former two ($\tilde{I}_b = \tilde{I}_e - \tilde{I}_c$). Since $\tilde{I}_c = \alpha \tilde{I}_e$, it follows that the *common emitter current amplification factor*

$$\beta = \frac{\tilde{I}_c}{\tilde{I}_b} = \frac{\tilde{I}_c}{\tilde{I}_e - \tilde{I}_c} = \frac{\alpha}{1 - \alpha}$$

can be made very high ($\sim 10^4$), but from considerations of stability and of frequency response, it is usually held in modern transistors in the range from 40 to 100.

Transistors have found universal application in electronics: in low- and high-frequency amplifier and oscillator circuits, in switching circuits, in triggers and multivibrators, in low- and high-frequency detector circuits, etc. Almost the whole of modern commercial and special-purpose electronics is based on semiconductor devices the most important of which is the transistor.

Photoelectric devices. p-n junction photocells. When a p-n junction is illuminated an emf is established in it. This phenomenon is utilized in barrier layer photocells which may serve as indicators of radiative energy independent of external power sources and as converters of radiative energy into electrical energy.

Figure 8.23(a) shows the schematic representation of a photocell. A narrow diffused n-layer is fabricated on the surface of a p-type semiconductor wafer so that a p-n junction is formed. In the absence of illumination, the p-n junction is in the state of equilibrium and an equilibrium potential barrier qV_c [Figure 8.23(b)] is established in it. When the junction is illuminated, electron-hole pairs are generated mostly in the p-region because light passes through the narrow n-layer without absorption. The electrons generated in the p-region diffuse to the p-n junction and are drawn in by the contact field and transported to the n-region. The holes are unable to surmount the barrier qV_c and remain in the p-region. Because of that,

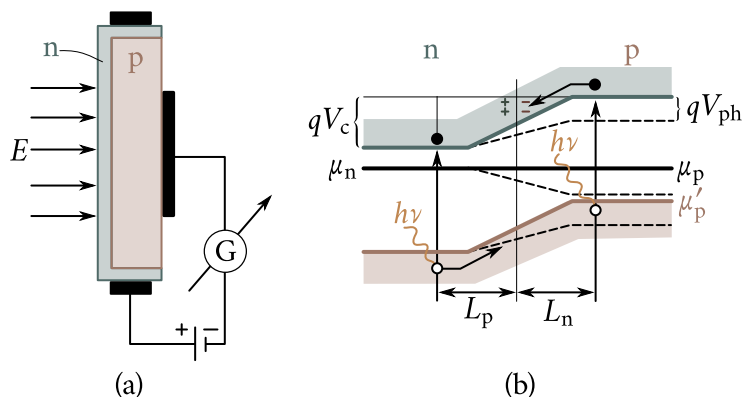


Figure 8.23: Semiconductor barrier layer photocell: (a)—schematic representation; (b)—energy band pattern of p-n junction.

the p-region acquires a positive charge and the n-region a negative one, and an additional forward voltage V_{ph} is established across the junction. The term for it is *photo-emf*, or *photovoltage*.

At present the most efficient converters of solar energy are silicon photocells (solar batteries). They are used as power supplies for receivers and transmitters installed on satellites and even on the ground. Calculations show the maximum efficiency (theoretical) of the silicon energy converters to be as high as 22%-23%. The efficiency of the best modern types is about 15%. Germanium, copper oxide, selenium, silver sulfide, sulfurous thallium and other semiconductor photodiodes are widely used as indicators of radiative energy. Their integral (in the entire spectrum) sensitivity is much higher (10^2 - 10^3 times) than that of the external photoeffect cells. Their main disadvantage is their great inertiality.

Photodiodes. The photodiode is a photocell connected into a circuit in-series with an external power supply [Figure 8.24(a)]. In the absence of illumination I , a negligible so-called *dark current* flows through the junction [Figure 8.24(b)]. When the p-n junction is illuminated, excess carriers are generated and the current rises in proportion to I causing a voltage drop across the load resistor R_L . Substantial advantages of the photodiodes over the external photoeffect elements are smaller dimensions and lower weight, high integral sensitivity and low operating voltage.

Luminescent diodes. The passage of a forward current through the p-n junction involves, as we know, minority carrier injection: of electrons into the p-region and of holes into the n-region. The injected carriers recombine with the majority carriers of the respective region, their intensity decreasing with the distance from the p-n junction [Figure 8.18(b)]. In many semiconductors, the recombination is nonradiative: the energy liberated in the recombination process is absorbed by the

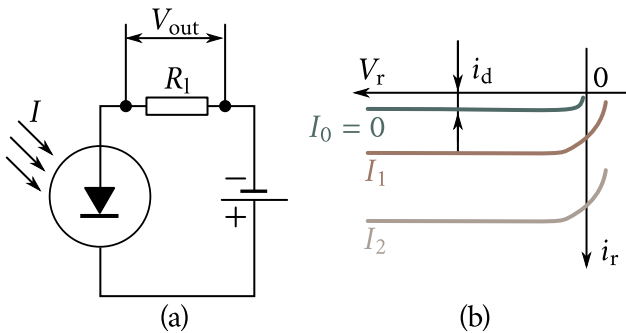


Figure 8.24: Photodiode: circuit diagram (a) and current-voltage characteristic (b).

crystal lattice, that is, turns eventually into heat. However, in such semiconductors as SiC, GaAs, InAs, GaP, and InSb, the recombination is radiative: the energy of recombination is liberated in the form of radiation quanta, photons.

Because of that a forward current flowing through the p-n junction made of such materials is accompanied by the emission of light from the junction region.

This phenomenon is utilized in the luminescent diodes. Such diodes are used in displays, they may be used in computers for data input and output and in other applications requiring reliable luminous indicators. Low operating voltages, low power consumption and a long service life are the advantages of the light emission diodes over other electroluminescent light sources.

Semiconductor lasers. In recent years, intensive work has been in progress on the semiconductor sources of coherent radiation—*i.e.*, semiconductor lasers—which open up possibilities for the direct conversion of electric energy into the energy of coherent radiation.

In Figure 8.25(a), the solid line shows the electron distribution corresponding to the equilibrium state and the dotted line the distribution corresponding to the nonequilibrium state, in which the concentrations of electrons in the conduction band and of the holes in the valence band are above the equilibrium values. Such band occupancy, corresponding to population inversion, is shown in Figure 8.25(b). The peculiar point about it is that the light quanta with the energy $\hbar\omega = E_g$ (E_g is the forbidden band width) cannot be absorbed by the system. Indeed, such an absorption involves the transfer of an electron from the top level of the valence band to the lowest level of the conduction band. Since there are practically no electrons on the top levels of the valence band and no vacant states at the bottom of the conduction band, the probability of such a process is extremely small. This creates favourable conditions for the stimulated emission and for an avalanche of photons. The light quantum 1 [Figure 8.25(b)] stimulates the recombination of the

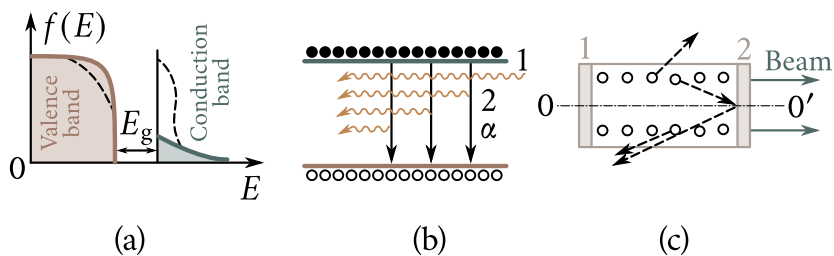


Figure 8.25: (a)—electron distribution plots for equilibrium (solid line) and inverse (dotted line) states of semiconductor; (b)—development of photon avalanche caused by induced radiation of a quantum system with population inversion; (c)—in a quantum generator only the radiation which propagates along the $00'$ axis is amplified.

electron and the hole (α -transition), which results in the emission of an identical quantum 2. Since the quanta are not absorbed by the system, they subsequently stimulate the emission of two new quanta, etc. To make one photon take part in many stimulated emission acts, two strictly parallel mirrors 1 and 2 [Figure 8.25(c)] are arranged on the opposite sides of the laser crystal to reflect the incident photons and return them into the crystal. Only those photons are amplified, which move strictly along the $00'$ axis, for only such photons are repeatedly reflected by mirrors 1 and 2. All other photons leave the active laser space immediately or after a limited number of reflections [in Figure 8.25(c) such photons are shown by dotted lines]. The result is a highly directional and highly monochromatic beam of radiation along the $00'$ axis.

There are various methods of creating a population inversion of the semiconductor energy bands. The best prospects offers the minority carrier injection through a forward biased p-n junction made in a degenerate semiconductor. Figure 8.26 shows the structure of a semiconductor laser in which such method of pumping is used. The laser is a diode with a p-n junction 1 made in the form of a bar. Highly polished faces 2 of this bar made strictly parallel to each other play the part of mirrors that reflect the photons.

The interest for the semiconductor lasers is due to some of their remarkable properties.

First of all, they have a high efficiency which may in principle reach 100%. This is due, on the one hand, to the quantum mechanical nature of the laser as a system in which only the “working” energy levels are excited, and on the other, to the fact that in a semiconductor laser, the electric energy is directly converted into coherent radiation without any intermediate steps as in all the other laser types.

Another remarkable property of the semiconductor laser is that it is possible to modulate the coherent radiation directly by changing the current through the

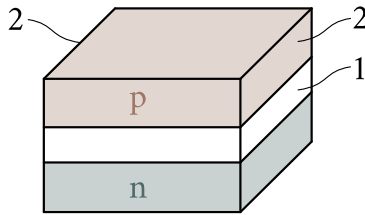


Figure 8.26: Schematic representation of semiconductor laser: 1—active region of laser; 2—reflecting faces (mirrors).

p-n junction. This enables them to be used in communications and in television as well as in ultra-high-speed computers for which their miniature dimensions are of special importance.

§ 78. Fundamentals of integrated circuit electronics (microelectronics)

The progress of modern science and technology requires electronics to provide for it efficient complex electronic equipment. Such equipment often contains hundreds of thousands of elements connected into a circuit by means of a similar number of connections.

Electronic equipment is responsible for the main part of costs of production of modern military equipment and aircraft. According to Western sources, the cost of electronic equipment makes up over 70% of the cost of a modern guided missile and over 50% of the cost of a modern bomber.

To produce such highly sophisticated equipment some difficult problems had to be solved, including the problems of drastic reduction in weight, dimensions, power consumption, and in price and of increasing the reliability of electronic devices.

Should modern electronic apparatus be assembled from components manufactured by the industry several decades ago, its weight would have been tons, its dimensions cubic metres, and its power consumption hundreds of kilowatts. The latter fact would have sufficed to make it impractical for many fields of the economy.

But those are not the only drawbacks. As the complexity of the electronic equipment increases its reliability—a factor of primary importance especially in military, computer and automatic production line applications—diminishes.

Finally, the problem of bringing down the costs of electronic equipment is also not devoid of importance and this problem can only be solved by far-reaching automation of the technological processes, which in turn requires the development of appropriate technology.

The efforts to solve those problems, lead to the creation of miniature electronic elements and blocks based on solid-state technology and to the microminiaturization of electronic equipment, resulting finally in the birth of a new field of modern electronics—microelectronics—whose main objective is the production of highly reliable and economical microminiature electronic circuits and apparatus.

The modern way to solve this problem is to devise new principles of constructing electronic circuitry, which would make possible the formation of a circuit as a whole on a miniature semiconductor crystal instead of assembling it from separate components. Such solid electronic circuits designed for specific applications are termed *integrated circuits* (IC). The integrated circuit, like an ordinary electronic circuit, is made up of active elements (transistors, diodes) and of passive elements (capacitors and resistors).

A semiconductor IC is fabricated on a single-crystal wafer (usually silicon) with the aid of methods of local doping with appropriate impurities to produce on it transistors, diodes, capacitors, and resistors and to connect them into a circuit. The dimensions of the wafer are typically $(10^{-2}) \times (5 \times 10^{-3})$ times $(2 \times 10^{-4}) \text{ m}^3$, the area of the active elements, for instance of a transistor, being under 10^{-9} m^2 .

The integrated circuits are usually characterized by packing density and by degree of integration. The packing density is the number of elements per unit volume of the IC, and the degree of integration, the number of elements making up the IC. Table 8.2 presents data on the packing density and on failure rate of circuits of different generations.

It follows then that, the changeover from the circuits assembled from pre-1941 components to modern IC increased the packing density some 10^5 times. There are reports of packing densities of IC of up to 10^{15} m^{-3} .

The degree of integration of an IC may vary in a wide range—from tens to hundreds or thousands of elements on each wafer. IC of over 100 elements are termed *big integrated circuits* (BIC).

The power consumption of IC, depending on the type, lies in the range of hun-

Table 8.2

Circuits using	Number of elements per m^3	Failure rate, $\lambda \text{ (h}^{-1}\text{)}$
Pre-1941 elements	3.5×10^4	10^{-5}
Miniature elements	1.8×10^5	5×10^{-6}
Semiconductor IC	3.0×10^9	Negligible

dreds of milliwatts to several microwatts.

Thus, the changeover to electronic equipment designed around IC practically solved the problems of dimensions, weight, and power consumption. Electronic computers are an impressive example of this. The first Soviet computers which were assembled from vacuum tubes and radio components (Minsk, Ural, etc.) occupied whole buildings, weighed tons, and consumed tens of kilowatts of power.

Electronic blocks assembled from IC have dimensions of the order of 10^{-2} m^3 and consume only hundreds of watts while special computers used, for instance, for launching and controlling missiles and spacecraft have dimensions of the order of 10^{-3} m^3 , weigh tens of kilograms and consume power of the order of tens of watts.

Presently BIC are being used in single wafer electronic calculators. The computer wafer (called "chip") of such a device is $(5 \times 5)10^{-6} \text{ m}^2$ and contains about 5000 transistors. BIC for electronic time-pieces including wrist watches have been developed as well. In such watches two wafers containing about 2000 transistors are used.

A substantial advantage of IC is that because mass-produced IC are much cheaper than equivalent circuits assembled from components. Modern technology makes it possible to arrange about a thousand IC on one single-crystal wafer of $5 \times 10^{-2} \text{ m}$ in diameter; if a hundred such wafers are processed at a time, about a million IC can be produced in one technological cycle.

The progress in microelectronics is a very rapid one. During the last decade a distance was covered from the simplest IC to BIC. In the nearest future, it is expected that most electronic equipment shall be based on integrated circuitry with the degree of integration increasing 100 to 1000-fold and a much greater reliability being attained. However, there are obstacles on this road, namely the so-called "tyranny of numbers" of microelements which already today crowd complex equipment in tens or hundreds of millions. To overcome this obstacle, it will, probably, be necessary to change over from the conventional IC to functional circuits, that is, to devices designed for specific functions and operating on some specific principle of solid state physics as a whole and not as a sum of individual elements (transistors, diodes, etc.).

As the simplest example of a functional device one may cite the ac-dc converter. The conventional circuit of such a converter consists of a transformer, rectifiers (semiconductor or vacuum diodes), and a filter. The functional converter consists of a resistance region in which the ac energy is transformed into heat, of the central low electric but high heat conductivity region, and of a thermoelectric region in which heat is converted into dc power. In such a device it is impossible to separate regions equivalent to the components of a conventional circuit. Here, the crystal

as a whole fulfills the complex function of an ac-dc converter.

The transition to functional circuits should result in a drastic decrease in the number of components and, therefore, in the decrease in the cost and in dimensions and in the improvement of reliability.

The process of creation of new scientific and technological trends in electronics and of devising devices and equipment based on new principles is a continuous one, the foundation for it being the utilization of the top-ranking achievements in the fundamental and applied sciences, first of all, in physics. Here, the leading role belongs to solid state physics which determines the mainstream of progress in modern electronics.

Chapter 9

Thermoelectric and Galvanomagnetic Phenomena

The thermoelectric effects include the Seebeck, the Peltier, and the Thomson effects, and the galvanomagnetic the Hall, the Ettingshausen, and the Nernst effects. Some of those phenomena have found wide application in practice; therefore, a look at them is not only of educational but of practical interest as well.

Let us discuss briefly the physical background of those phenomena.

§ 79. The Seebeck effect

In 1822, T. J. Seebeck discovered that an electromotive force V_T is established in a circuit consisting of two conductors 1 and 2 made of different materials if the junctions of these conductors, A and B, are kept at different temperatures, T_1 and T_2 [Figure 9.1(a)]. This emf is termed *thermal emf*. Experiments show it to be—in a narrow temperature interval—proportional to the difference in the temperature of the junctions A and B:

$$V_T = \alpha(T_2 - T_1). \quad (9.1)$$

The proportionality factor

$$\alpha = \frac{dV_T}{dT} \quad (9.2)$$

is called *differential*, or *specific, thermoelectric power*. It is determined by the material of the conductors and the temperature.

There are three sources of thermal emf: the directional current of the carriers in the conductor due to the presence of a temperature gradient (the volumetric component V_V), the change in the position of the Fermi level (the junction component V_j), and the drag of the electrons by the phonons (the so-called phonon drag

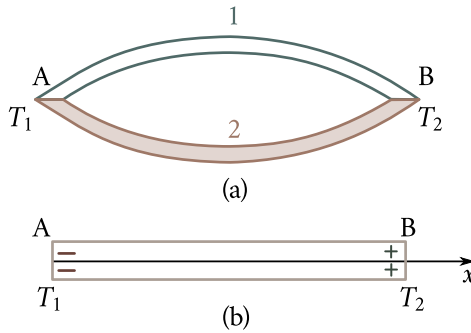


Figure 9.1: The Seebeck effect: (a)—thermoelectric circuit; (b)—origin of volumetric and junction components of thermal emf.

effect).

Let us discuss the physical nature of each of those sources.

Volumetric component of thermal emf. Suppose that a temperature difference $(T_2 - T_1)$ is maintained at the terminals of a uniform conductor AB [Figure 9.1(b)], so that there is a temperature gradient dT/dx in the direction from B to A. The current carriers in the hot end have greater kinetic energy and greater speeds of motion than the carriers in the cold end. Therefore, a current will flow in the conductor from the hot end to the cold; this current will charge the conductor. In cases when the current is carried by electrons, the cold end will accumulate a negative charge and the hot end a positive charge, and a potential difference V_V will be established between them. This is the volumetric component of thermal emf. The differential thermoelectric power corresponding to this component is

$$\alpha_V = \frac{\partial V_V}{\partial T}, \quad (9.3)$$

α_V may be estimated as follows. The pressure of the electron gas in the conductor is

$$p = \frac{2}{3} n \bar{E}, \quad (9.4)$$

where \bar{E} is the average energy of electrons in the conductor, and n their concentration.

The temperature gradient occasions a pressure gradient to compensate which a field \mathcal{E} should be established in the conductor such that

$$q\mathcal{E}n = \frac{\partial p}{\partial x} = \frac{\partial p}{\partial T} \frac{\partial T}{\partial x}.$$

From here, α_V may easily be found:

$$\alpha_V = \frac{\partial V_V}{\partial T} = \mathcal{E} \left(\frac{\partial T}{\partial x} \right)^{-1} = \frac{1}{nq} \frac{\partial p}{\partial T}. \quad (9.5)$$

As a rule, in an n-type conductor, α_V is directed from the hot end to the cold. However, there are exceptions to this rule which we will discuss below.

The junction component of thermal emf. The change in temperature occasions a change in the position of the Fermi level. In n-type conductors, the Fermi level sinks on the energy diagram as the temperature is raised [see Figure 5.19(a)]. By force of this it should be higher on the cold end of an n-type conductor than on its hot end. The difference in the Fermi level positions is equivalent to a potential difference

$$dV_j = -\frac{1}{q} \frac{\partial \mu}{\partial T} dT. \quad (9.6)$$

And this is just the *junction component* of the thermal emf. The differential thermoelectric power corresponding to this component is

$$\alpha_j = -\frac{1}{q} \frac{\partial \mu}{\partial T}. \quad (9.7)$$

The resultant differential thermoelectric power

$$\alpha = \frac{1}{nq} \frac{\partial p}{\partial T} - \frac{1}{q} \frac{\partial \mu}{\partial T}. \quad (9.8)$$

We apply the relation (9.8) to conductors of various kinds.

Thermoelectric power of metals. Substituting the average energy of electrons of a degenerate electron gas from Eq. (3.45) into (9.4), we obtain the following expression for the pressure of the electron gas in a metal:

$$p = \frac{2}{3} n \bar{E} = \frac{2}{5} n E_F + \frac{\pi^2}{6 E_F} (k_B T)^2.$$

Differentiating this expression with respect to T and multiplying it by $1/nq$, we obtain

$$\alpha_V = \frac{k_B}{q} \frac{\pi^2}{3} \frac{k_B T}{E_F}. \quad (9.9)$$

The temperature dependence of the Fermi level in metals is given by relation (3.44):

$$\mu = E_F \left[1 - \frac{\pi^2}{12} \left(\frac{k_B T}{E_F} \right)^2 \right].$$

Differentiating it with respect to T and multiplying by $1/q$, we obtain

$$\alpha_j = -\frac{\pi^2 k_B}{6q} \frac{k_B T}{E_F}. \quad (9.10)$$

Substituting Eqs. (9.9) and (9.10) into (9.8), we obtain

$$\alpha_m = \frac{\pi^2 k_B}{3q} \left(1 + \frac{1}{2} \right) \frac{k_B T}{E_F}. \quad (9.11)$$

A more accurate calculation for metals with a quadratic dependence of the electron energy on the wave vector produces the following result:

$$\alpha_m = \frac{\pi^2 k_B}{3q} (1 + r) \frac{k_B T}{E_F}, \quad (9.12)$$

where r is the exponent in the relation

$$\lambda \propto E^r, \quad (9.13)$$

which expresses the dependence of the electron mean free path on energy E .

Table 9.1 presents the values of r for various mechanisms of electron scattering on atomic-and ionic lattices.

It follows from Eq. (9.12) that for metals $\alpha_m \propto T$, in full agreement with experiment. Since $k_B T \ll E_F$, the thermoelectric power of metals is quite small. For instance, for silver $E_F = 5.5$ eV and $k_B T = 0.025$ eV at $T = 300$ K. Substituting this into Eq. (9.12), we obtain $\alpha_m \approx 8 \times 10^{-6}$ V K⁻¹, which is quite close to the experimental value $\alpha_m \approx 5 \times 10^{-6}$ V K⁻¹.

It follows from Eq. (9.13), that when $r < 0$ the more energetic electrons have a shorter mean free path λ . Since the diffusion current in this case is directed from the hot end to the cold end, the sign of the volumetric component of the thermal emf will be reversed. This may cause the reversal of the sign of the metal's thermal emf as a whole. Such effects are observed, for example, in some transition metals and in some alloys.

As was already stated before, Eq. (9.12) is valid for metals with a quadratic $E(k)$ dependence. In metals and in alloys with a complex Fermi surface, the contribution of the various regions of this surface may differ not only in absolute value but in sign as well, with the result that the thermoelectric power may be zero or close to zero. For instance, lead has a zero thermoelectric power. For this reason, thermoelectric power is usually measured in relation to lead.

The current direction in the hot junction of a thermocouple made of an n-type conductor and of lead will be determined by the polarity of the conductor's

Table 9.1

Scattering on thermal vibrations				Scattering on impurity atoms
Atomic lattice	Ionic lattice			
	$T < \theta$	$T > \theta$		
r	0	1/2	1	2

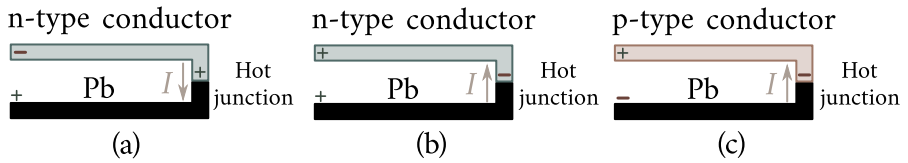


Figure 9.2: Magnitude and sign of thermoelectric power determined in relation to lead (explanation in text).

charge. For a normal conductor whose hot junction acquires a positive charge, the current in it will be directed from conductor to lead [Figure 9.2(a)]. In this case, the conductor's thermoelectric power is assumed to be negative.

In case of an n-type conductor acquiring an anomalous charge [Figure 9.2(b)], the current in the hot junction will flow from lead to conductor and it will be positive; λ will be positive for a normal p-type conductor too, whose hot end acquires a negative charge [Figure 9.2(c)].

Thermoelectric power of semiconductors. The pressure of electron gas in a nondegenerate semiconductor is

$$p = \frac{2}{3} n \bar{E} = nk_B T.$$

Differentiating this expression with respect to T and multiplying by $1/nq$, we obtain

$$\alpha_V = \frac{k_B}{q} \left[1 + T \frac{\partial \ln n}{\partial T} \right]. \quad (9.14)$$

A more rigorous calculation yields

$$\alpha_V = \frac{k_B}{q} \left[r + \frac{1}{2} + T \frac{\partial \ln n}{\partial T} \right]. \quad (9.15)$$

The chemical potential in a nondegenerate n-type semiconductor is given by relation (3.26):

$$\mu_n = k_B T \ln \left[\frac{nh^3}{2(2\pi m_n k_B T)^{3/2}} \right].$$

Differentiating μ_n with respect to T and multiplying by $1/q$, we obtain

$$\alpha_j = \frac{k_B}{q} \left(\frac{3}{2} - \frac{\mu_n}{k_B T} - T \frac{\partial \ln n}{\partial T} \right). \quad (9.16)$$

Substituting Eqs. (9.15) and (9.16) into (9.8), we obtain

$$\alpha_n = -\frac{k_B}{q} \left(r + 2 - \frac{\mu_n}{k_B T} \right) = -\frac{k_B}{q} \left\{ r + 2 + \ln \left[\frac{2(2\pi m_n k_B T)^{3/2}}{nh^3} \right] \right\}. \quad (9.17)$$

The minus sign in front of the right-hand side is in accordance with the conventional polarity of the thermoelectric power.

For a n-type semiconductor

$$\alpha_p = \frac{k_B}{q} \left\{ r + 2 + \ln \left[\frac{2(2\pi m_p k_B T)^{3/2}}{ph^3} \right] \right\}. \quad (9.18)$$

Let us estimate the value of α of an extrinsic semiconductor, for instance, for n-type germanium with $n = 10^{23} \text{ m}^{-3}$ at $T = 300 \text{ K}$. Substituting those values into Eq. (9.17), we obtain $\alpha \approx 10^{-3} \text{ V K}^{-1}$. Hence, the thermoelectric power of semiconductors is three orders of magnitude greater than that of metals.

For semiconductors with bipolar conductivity, in which the electric current is carried both by electrons and holes, the expression for the thermoelectric power is:

$$\alpha_{n,p} = \frac{\alpha_p p u_p - \alpha_n n u_n}{p u_p + n u_n}. \quad (9.19)$$

It follows from this relation that if the electron and hole concentrations and their mobilities turn out to be equal, the thermoelectric power may be quite small or even zero.

Phonon drag of electrons. The phonon drag effect, discovered by L. E. Gurevich in 1945, consists in the following. With a temperature gradient in the conductor, the phonons drift from its hot end to the cold end at an average velocity v_{ph} . In the presence of such drift, the electrons scattered by the drifting phonons are themselves involved in the directional motion from the hot end to the cold end, their velocity being about equal to v_{ph} . The accumulation of the electrons on the cold end of the conductor and their depletion on the hot end results in the appearance of a thermal emf V_{ph} .

G. E. Pikus in 1956 calculated the differential thermoelectric power due to the phonon drag and obtained the following result:

$$\alpha_{ph} = \frac{k_B}{3q} \frac{m_n v_{ph}^2}{k_B T} \frac{\tau_{ph}}{\tau_e}. \quad (9.20)$$

Here, v_{ph} is the phonon drift velocity, and τ_{ph} and τ_e are the phonon and electron relaxation times, respectively.

In the low temperature range, this component of thermoelectric power can be tens or hundreds of times greater than the volumetric and junction components.

§ 80. The Peltier effect

Let a current I flow in a circuit consisting of two conductors 1 and 2 (Figure 9.3) made of different materials. The Joule heat $Q = I^2 R t$ will be liberated in the junctions A and B (R is the junction's resistance, and t is the time the current flows).

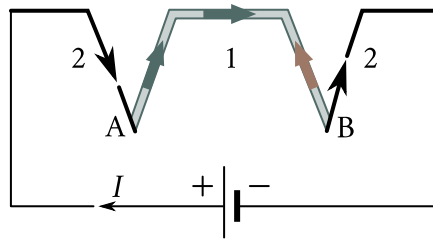


Figure 9.3: Diagram of circuit used to observe the Peltier effect.

At junctions of identical conductors only this heat will be liberated and, from this point of view, there is no difference between the junction and the rest of the circuit. At the same time, at points of junction of different materials, an additional heat apart from the Joule heat will be liberated or absorbed, heating the contact in the former case or cooling it in the latter. This phenomenon was discovered in 1834 by J. C. A. Peltier and is termed the *Peltier effect*; the additional heat liberated or absorbed in the junction is termed *Peltier heat*, Q_P . Experiments show it to be proportional to the current I and the time the current passes through the contact t :

$$Q_P = \Pi It. \quad (9.21)$$

The proportionality factor Π is termed the *Peltier coefficient*. Its value is determined by the materials and by temperature.

There is a direct connection between the Peltier and Seebeck effects: the temperature difference causes a current to flow in a circuit consisting of different materials and a current flowing through such a circuit sets up a temperature difference. The expression for this relation is due to W. Thomson (Lord Kelvin), who is the author of the thermodynamic theory of thermoelectric phenomena. He demonstrated that:

$$\alpha = \frac{\Pi}{T}. \quad (9.22)$$

The Peltier effect is due to the difference in the average energies of the conduction electrons in unlike materials. By way of an example, let's consider the junction of a metal with a nondegenerate n-type semiconductor (Figure 9.4). After equilibrium had been established their Fermi levels will coincide. Only the electrons close to the Fermi level whose average energy are practically equal to the Fermi energy, take part in the conductivity in the metal.

Denote the average energy of the conduction electrons in the semiconductor by \bar{E}_n . This energy is not equal to the thermal energy of the electrons $3k_B T/2$ since the relative part played by fast electrons in the electric current is greater than that

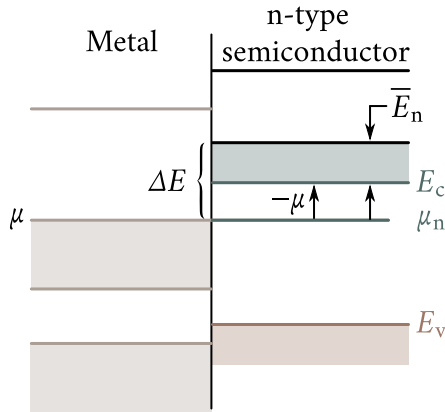


Figure 9.4: Energy band pattern of metal-semiconductor junction illustrating mechanism of the Peltier effect.

of slow electrons. A calculation made for the case of a nondegenerate electron gas yields

$$\bar{E}_n = (r + 2)k_B T, \quad (9.23)$$

where r is the exponent in Eq. (9.13).

Suppose that the electric current flowing through the junction is such that the electrons flow from the semiconductor to the metal. It may be seen from Figure 9.4, that each electron that goes over from the semiconductor to the metal carries with it an additional energy equal to

$$\Delta E = \bar{E}_n + (-\mu_n). \quad (9.24)$$

This energy is the Peltier heat and it is liberated near the junction. When the direction of the current is changed, the electrons going over from the metal to the semiconductor absorb heat and cool the junction.

Dividing ΔE by the electron charge, we obtain the Peltier coefficient:

$$\Pi_{mn} = -\frac{\Delta E}{q} = -\frac{1}{q}(\bar{E}_n - \mu_n). \quad (9.25)$$

Substituting μ_n from Eq. (3.26) and \bar{E}_n from Eq. (9.23) into (9.25), we obtain

$$\Pi_{mn} = -\frac{k_B T}{q} \left\{ (r + 2) + \ln \left[\frac{2(2\pi m_n k_B T)^{3/2}}{n h^3} \right] \right\}. \quad (9.26)$$

A similar relation may be obtained for the junction of a metal with a p-type semiconductor:

$$\Pi_{mp} = -\frac{k_B T}{q} \left\{ (r + 2) + \ln \left[\frac{2(2\pi m_p k_B T)^{3/2}}{p h^3} \right] \right\}. \quad (9.27)$$

For a junction of two metals the Peltier coefficient may be found from Eq. (9.22):

$$\Pi_{1,2} = (\alpha_1 - \alpha_2)T. \quad (9.28)$$

Substituting α from Eq. (9.12), we obtain

$$\Pi_{1,2} = \frac{\pi^2 k_B^2 T^2}{3q} (1 + r) \left(\frac{1}{E_{F1}} - \frac{1}{E_{F2}} \right). \quad (9.29)$$

§ 81. The Thomson effect

Imagine a homogeneous conductor AB with a temperature gradient dT/dx along its length and carrying a current I [Figure 9.1(b)]. W. Thomson predicted theoretically that in such a conductor, apart from the Joule heat, an additional amount of heat Q_τ proportional to the current I , the temperature difference $(T_2 - T_1)$, and the time t should be liberated or absorbed depending on the direction of the current:

$$Q_\tau = \tau I (T_2 - T_1) t. \quad (9.30)$$

The heat Q_τ is termed the *Thomson heat* and the proportionality factor τ the *Thomson coefficient*. It is determined by the material of the conductor and by temperature. According to Thomson's theory, the difference in Thomson coefficients of two conductors is related to their differential thermoelectric power by the expression:

$$\frac{d\alpha_{1,2}}{dT} = \frac{\tau_1 - \tau_2}{T}. \quad (9.31)$$

The Thomson effect is due to the fact that in a conductor in which a temperature gradient exists, the carrier flux carries not only the electric charge but heat as well. Suppose the current in the conductor AB [Figure 9.1(b)] flows in the direction corresponding to the electron flow from the hot end B to the cold end A. The "hot" electrons as they arrive in the cold regions give up their extra energy and heat the conductor. When the direction of the current is changed, the conductor is cooled.

In quantitative calculations of the Thomson effect one should take into account the thermal emf set up in the conductor, which in the former case will retard the electrons and in the latter accelerate them. This thermal emf can change not only the magnitude of the Thomson coefficient but even its sign.

§ 82. Galvanomagnetic phenomena

The Hall effect. Suppose a current of density i flows in a conducting bar of width a and thickness b (Figure 9.5). Choose points C and D on the side faces of the bar such that the potential difference between them is zero. Should this bar be placed into

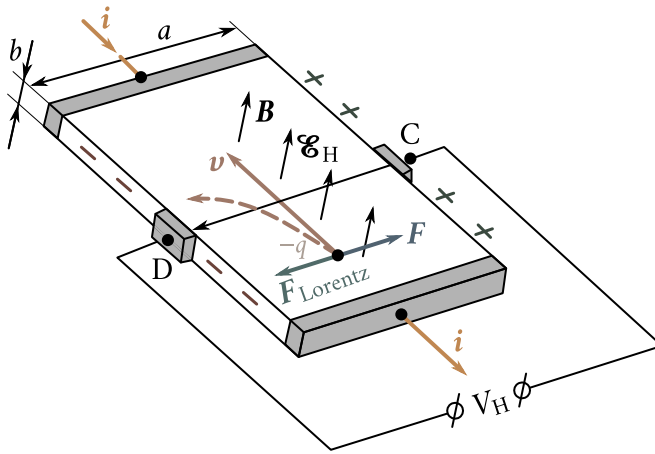


Figure 9.5: Layout used to observe the Hall effect.

a magnetic field with induction \mathbf{B} , potential difference V_H termed *Hall emf* would appear between points C and D. It follows from experiments that in magnetic fields not too strong

$$V_H = R_H B i a. \quad (9.32)$$

The proportionality factor R_H is termed the *Hall coefficient*. Its dimensions are $L^3 Q^{-1}$ (L is length and Q electric charge) and it is measured in cubic metres per coulomb ($\text{m}^3 \text{C}^{-1}$). Let us consider the physical origin of the Hall effect.

The Lorentz force $\mathbf{F}_{\text{Lorentz}}$ acting on an electron moving from right to left at a speed \mathbf{v} (Figure 9.5) is

$$\mathbf{F}_{\text{Lorentz}} = q\mathbf{v} \times \mathbf{B}.$$

If $\mathbf{v} \perp \mathbf{B}$, the force will be equal to

$$F_{\text{Lorentz}} = qvB.$$

The Lorentz force deflects the electrons to the outer face of the bar (dotted line in Figure 9.5), and the bar receives a negative charge. Uncompensated positive charges accumulate on the opposite side. This results in an electric field directed from C to D:

$$\mathcal{E}_H = \frac{V_H}{a},$$

where V_H is the potential difference between points C and D (the Hall emf).

The field \mathcal{E}_H exerts a force $\mathbf{F} = q\mathcal{E}_H$ on the electrons, this force being directed against the Lorentz force. When $F = F_{\text{Lorentz}}$, the transverse electric field compensates the Lorentz force and no more electric charges are accumulated on the side

faces of the bar. From the conditions of equilibrium

$$qvB = q\mathcal{E}_H, \quad (9.33)$$

we obtain

$$\mathcal{E}_H = vB.$$

Multiplying this relation by the distance a between points C and D, we obtain

$$V_H = a\mathcal{E}_H = vBa.$$

Taking into account that $i = nqv$ and, consequently $v = i/nq$, we obtain

$$V_H = \frac{1}{nq} Bia. \quad (9.34)$$

Thus, theory produces an expression for V_H that coincides with the relation Eq. (9.32) obtained from experiment. The Hall constant turns out to be equal to

$$R_H = \frac{1}{nq}. \quad (9.35)$$

It follows from Eq. (9.35) that, knowing the absolute value and the sign of the Hall constant, we can find the concentration and sign of the charge carriers in a conductor; R_H of n-type conductors is negative and of p-type conductors positive.

If we measure in addition the specific conductance $\sigma = nqu$ of the conductor, we will be able to find the carrier mobility u since

$$R_H\sigma = u_H. \quad (9.36)$$

Mobility u_H determined from Eq. (9.36) is the *Hall mobility* and it may not coincide with the drift mobility defined by Eq. (6.5).

In the derivation of Eq. (9.35), it was assumed that all carriers in the conductor have the same speed v . Such an assumption is valid in case of metals and degenerate semiconductors but it is totally unacceptable for nondegenerate semiconductors, in which the carrier velocities are distributed in accordance with the Boltzmann law. A more rigorous derivation, which accounts for this fact, yields the following expression for R_H :

$$R_H = \frac{A}{nq} \quad (9.37)$$

where A is a constant dependent on the scattering mechanism of carriers in the crystal. The typical values of A are given in Table 9.2 below.

In bipolar semiconductors the current is carried simultaneously by electrons and holes. Since their charges are opposite and they move in opposite directions in an electric field, the Lorentz force $\mathbf{F}_{\text{Lorentz}} = q\mathbf{v} \times \mathbf{B}$ deflects them in the same direction. Because of this, other conditions equal, their Hall emf and Hall coefficients will be smaller than in unipolar semiconductors. Calculation yields the following

expression for R_H of bipolar semiconductors:

$$R_H = \frac{A}{q} \left[\frac{u_p^2 p - u_n^2 n}{(u_p + u_n)^2} \right] \tag{9.38}$$

where n and p are electron and hole concentrations, and u_n and u_p their mobilities. Depending on which of the two terms in the numerator is greater, the sign of the Hall coefficient may either be positive, or negative.

For intrinsic semiconductors, in which $n = n_i$, Eq. (9.38) assumes the form:

$$R_H = \frac{A}{n_i q} \left(\frac{u_p - u_n}{u_p + u_n} \right). \tag{9.39}$$

It follows from Eq. (9.39) that, in the intrinsic range, the sign of the Hall coefficient is determined by that of the carriers with greater mobility. As a rule such carriers are electrons. Therefore, when an extrinsic p-type semiconductor goes over to the intrinsic range, the sign of the Hall coefficients changes. Hall coefficient (at room temperature) for some metals and intrinsic semiconductors is presented below in Table 9.3.

As indicated, the Hall coefficient of semiconductors is many orders of magnitude greater than that of metals. The explanation is that the carrier concentration in semiconductors is much less but the mobility, on the other hand, is much greater than in metals.

Table 9.2

Scattering on thermal vibrations				
Atomic lattice		Ionic lattice		Scattering on impurity atoms
		$T < \theta$	$T > \theta$	
A	1.17	0.99	1.11	1.93

Table 9.3

	Cu	Zn	Bi	Ge	Si
$R_H, (10^{-11} \text{ m}^3 \text{ C}^{-1})$	5.5	3.3	10^3	10^{10}	10^{13}

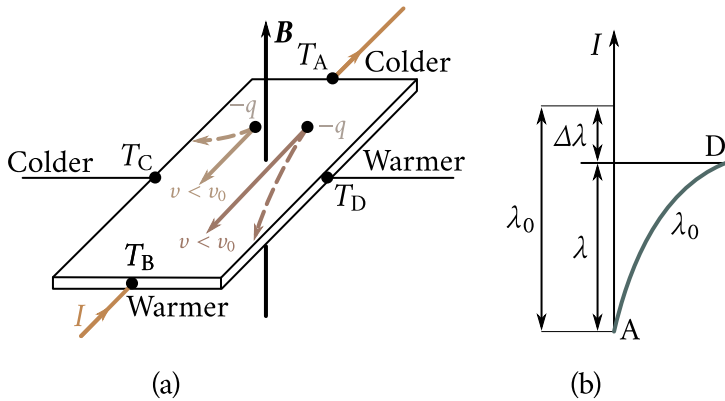


Figure 9.6: Diagram explaining the origin of the Ettingshausen and Nernst effects (a) and of specific resistance variations in magnetic field (b).

Ettingshausen effect. The thermal velocities of electrons in nondegenerate semiconductors lie in a wide range. In such conditions, Eq. (9.33) may be valid not for all the electrons but only for some of them whose average velocity is v_0 . For electrons whose velocity is $v > v_0$, we have $qvB > q\mathcal{E}_H$ and they will be deflected to the right-hand face of the plate [Figure 9.6(a)]. The electrons whose velocity is $v < v_0$, so that $qvB < q\mathcal{E}_H$, will be deflected to the left-hand face of the plate.

Fast electrons reaching the right-hand face give up their extra energy to it and thereby heat it. The slow electrons, which accumulate on the left-hand face, replenish their energy deficit at the expense of the thermal energy of the crystal and thereby cool it. Thus, a transverse temperature difference $T = T_D - T_C$ is established. This phenomenon is termed *Ettingshausen effect*.

Nernst effect. The electrons entering a homogeneous magnetic field B perpendicular to their velocity v start moving in a circle with a radius

$$r = \frac{m_n v}{qB}. \quad (9.40)$$

It follows from Eq. (9.40) that the fast electrons are rotated by the magnetic field less than the slow electrons [Figure 9.6(a)]. Therefore, the front face of the plate will be richer in hot electrons and will be heated while the back face will be richer in slow electrons and will be cooled. A longitudinal temperature difference $T_B - T_A$ will be established. This is the *Nernst effect*.

Variation of conductor's resistance in magnetic fields (magnetoresistance). The trajectories of electrons moving in a magnetic field with velocities other than v_0 are curved [Figure 9.6(a)] and this results in a reduction of their effective mean free path in the direction of the electric current. If the mean free path in the direction of the current in the absence of a magnetic field is λ_0 , in the magnetic

field it is equal to the projection of the arc AD on the direction of the current I , that is $\lambda = \lambda_0 - \Delta\lambda$ [Figure 9.6(b)].

Since the carrier mobility u is proportional to the mean free path, the decrease in the mean free path by $\Delta\lambda$ should bring about a decrease in mobility by Δu and in the semiconductor's conductivity by $\Delta\sigma$, so that

$$\frac{\Delta\sigma}{\sigma} = \frac{\Delta u}{u} = \frac{\Delta\lambda}{\lambda_0}.$$

The theory provides the following expression for the relative increase in specific resistance of extrinsic unipolar semiconductors:

$$\frac{\Delta\rho}{\rho} = cu^2 B^2, \quad (9.41)$$

where B is the magnetic field induction, and c a coefficient dependent on the carrier scattering mechanism.

The ratio $\Delta\rho/\rho$ is termed *magnetoresistivity*. It follows from Eq. (9.41) that, by measuring magnetoresistivity, one can directly find the carrier mobility.

§ 83. Practical applications of thermoelectric and galvanomagnetic phenomena

Thermoelectric phenomena. For a long time the only application of the Seebeck effect was in measurements. Placing one junction of a thermocouple in a thermostat, held at a known constant temperature and the other junction into the object the temperature of which is to be measured, one can determine this temperature from the thermal emf established in the couple. Such measurements are quite simple, reliable and sufficiently accurate and can be carried out in a wide temperature range.

However, after semiconductors were discovered it became possible to use the Seebeck effect for the direct conversion of thermal energy into electric energy. The devices used to this end are termed thermoelectric generators and the elements of which they are assembled thermoelements. The man mainly responsible for their development and wide publicity was the Soviet physicist A. F. Ioffe.

The first thermoelectric generators were produced before World War II and were used in the war to power radio equipment. The thermoelectric generators were mounted in the bottom of a special kettle and heated in the process of boiling water.

In 1953 a commercial type of thermoelectric generator of 3 W power for battery receivers was produced; later thermoelectric generators of 1 kW power and more were produced. Presently generators designed for hundreds of kilowatts are being developed.

The midseventies saw the appearance of thermoelectric generators utilizing the heat of radioactive decay of chemical elements. An example of such a generator is the generator Beta-1 with a power of 150 W to 200 W, which operates on the radioactive cerium isotope Ce-144. It was designed to power electronic equipment of automatic radiometeorologic stations, earth satellites, etc.

In 1964 an experimental atomic reactor-energy converter Romashka (Camomile) with a power of 500 W for direct conversion of heat energy into electric energy was built.

Work is in progress on thermoelectric generators that would utilize the thermal energy of the sun's radiation.

It is a regretful fact, but the efficiency of even the best modern experimental thermoelements does not rise above 8%.

The Peltier effect is beginning to be widely used in practice mainly for various cooling devices: home refrigerators, devices for cooling aircraft electronic equipment, microcoolers for biological applications, various thermoelectric thermostats, temperature-controlled microscope supports, etc. Quite possible, the Peltier effect will in the near future be used for heating dwellings in winter and for cooling them in summer.

Galvanomagnetic phenomena. The most widely used galvanomagnetic phenomenon is the Hall effect. Apart from applications in the study of electric properties of materials it served as a basis for the design of a wide class of instruments: magnetometers, dc-ac and ac-dc converters, signal generators, phase meters, microphones, etc.

In recent years attempts are made to use the Ettingshausen effect in cooling devices. With the right choice of materials and with the optimal geometry of the cooling crystal it is possible to obtain temperatures of the cold face of the crystal of over 100 °C below that of the surroundings.

APPENDICES

A.1. Derivation of the Maxwell-Boltzmann distribution function

To obtain expression (3.25), consider a collision of two particles one of which is in a state with the energy E_1 and the other in a state with the energy E_2 . After the collision the particles will go over to states with energies E_3 and E_4 , respectively. Let us define the term *reverse collision* as a collision that returns the particles to the initial states with the energies E_1 and E_2 . Thus, we shall consider collisions of two types:

$$(E_1, E_2) \rightarrow (E_3, E_4) \quad (\text{direct})$$

$$(E_3, E_4) \rightarrow (E_1, E_2). \quad (\text{inverse})$$

The rate of direct collisions Q_d is proportional to the average number of particles in the initial state, that is $f(E_1)$ and $f(E_2)$, and is independent of the number of particles in the final state because the gas is nondegenerate:

$$Q_d = cf(E_1)f(E_2), \tag{A.1}$$

where c is a proportionality factor.

The number of reverse collisions is proportional to $f(E_3)f(E_4)$:

$$Q_r = cf(E_3)f(E_4). \tag{A.2}$$

In the state of thermodynamic equilibrium Q_d should be equal to Q_r :

$$f(E_1)f(E_2) = f(E_3)f(E_4). \tag{A.3}$$

Making use of the energy conservation law, $E_1 + E_2 = E_3 + E_4$, we may rewrite the expression in the form:

$$f(E_1)f(E_2) = f(E_3)f(E_1 + E_2 - E_3). \tag{A.4}$$

Note that E_1, E_2, E_3 must be regarded as independent quantities. Taking the logarithm of both sides of Eq. (A.4), we obtain

$$\ln [f(E_1)f(E_2)] = \ln [f(E_3)f(E_1 + E_2 - E_3)]. \tag{A.5}$$

Differentiate this sum with respect to E_1 . Since E_2 and E_3 are independent of

E_1 , they may be assumed to be constant. Then,

$$\frac{1}{f(E_1)} \frac{df(E_1)}{dE_1} = \frac{1}{f(E_1 + E_2 - E_3)} \frac{df(E_1 + E_2 - E_3)}{dE_1} \frac{df(E_1 + E_2 - E_3)}{dE_1}. \quad (\text{A.6})$$

Since $df(E_1 + E_2 - E_3)/dE_1 = 1$, it follows that

$$\frac{1}{f(E_1)} \frac{df(E_1)}{dE_1} = \frac{1}{f(E_1 + E_2 - E_3)} \frac{df(E_1 + E_2 - E_3)}{dE_1}. \quad (\text{A.7})$$

Differentiate Eq. (A.5) with respect to E_2 :

$$\frac{1}{f(E_2)} \frac{df(E_2)}{dE_2} = \frac{1}{f(E_1 + E_2 - E_3)} \frac{df(E_1 + E_2 - E_3)}{dE_2}. \quad (\text{A.8})$$

Comparing Eq. (A.7) with Eq. (A.8), we obtain

$$\frac{1}{f(E_1)} \frac{df(E_1)}{dE_1} = \frac{1}{f(E_2)} \frac{df(E_2)}{dE_2}. \quad (\text{A.9})$$

The left-hand side of Eq. (A.9) is independent of E_2 , the right-hand side is independent of E_1 ; therefore, each of them is equal to some constant independent of the particles' energy. Denote it by β . Then, we may rewrite Eq. (A.9) as follows:

$$\frac{1}{f(E)} \frac{df(E)}{dE} = \beta. \quad (\text{A.10})$$

Integrating Eq. (A.10), we obtain

$$f(E) = A e^{\beta E}, \quad (\text{A.11})$$

where A is the integration constant. Experiment shows that

$$\beta = -\frac{1}{k_B T}, \quad A e^{\mu/(k_B T)}. \quad (\text{A.12})$$

Substituting Eq. (A.12) into Eq. (A.11), we finally get

$$f_M(E) = e^{\mu/(k_B T)} e^{-E/(k_B T)}. \quad (\text{A.13})$$

A.2. Derivation of the Fermi-Dirac distribution function

Consider, as we did in Appendix A.1, the direct and reverse particle collisions. Recall that in the case of a nondegenerate gas, the rate of collisions was independent of the number of particles in the final stages and was entirely determined by the number of particles in the initial stages. The situation in the case of a degenerate fermion gas is a different one: if a state is already occupied, it cannot accept another fermion and the collision will not take place. For this reason, in the case of a degenerate fermion gas, the rate of collisions is proportional not only to the average number of particles in the initial states but to the average number of vacant states with the energies E_3 and E_4 as well.

Since $f_F(E)$ expresses the probability for the state with the energy E to be occu-

pied, $1 - f_F(E)$ expresses the probability for it to be vacant. Therefore, the average numbers of vacant states with the energies E_3 and E_k are $1 - f_F(E_3)$ and $1 - f_F(E_4)$, respectively. Accordingly, the rates of direct and reverse collisions are:

$$Q_d = c f_F(E_1) f_F(E_2) [1 - f_F(E_3)] [1 - f_F(E_4)], \quad (\text{A.14})$$

$$Q_r = c f_F(E_3) f_F(E_4) [1 - f_F(E_1)] [1 - f_F(E_2)]. \quad (\text{A.15})$$

In the state of thermal equilibrium

$$\begin{aligned} f_F(E_1) f_F(E_2) [1 - f_F(E_3)] [1 - f_F(E_4)] \\ = f_F(E_3) f_F(E_4) [1 - f_F(E_1)] [1 - f_F(E_2)]. \end{aligned} \quad (\text{A.16})$$

Dividing both sides of Eq. (A.16) by $f_F(E_1) f_F(E_2) f_F(E_3) f_F(E_4)$, we obtain

$$\left[\frac{1}{f_F(E_1)} - 1 \right] \left[\frac{1}{f_F(E_2)} - 1 \right] = \left[\frac{1}{f_F(E_3)} - 1 \right] \left[\frac{1}{f_F(E_4)} - 1 \right]. \quad (\text{A.17})$$

Comparing Eq. (A.17) with Eq. (A.4), one may easily see that the function $1/f_F(E) - 1$, for a degenerate fermion gas, satisfies the same condition as is satisfied by the function $f_F(E)$ in the case of a nondegenerate gas. This makes it possible to use the result Eq. (A.10), which in this case takes the form

$$\left[\frac{1}{f_F(E)} - 1 \right]^{-1} \frac{d}{dE} \left[\frac{1}{f_F(E)} - 1 \right] = \gamma, \quad (\text{A.18})$$

where γ is a constant. Integrating Eq. (A.18), we obtain

$$\frac{1}{f_F(E)} = B e^{\gamma E}, \quad (\text{A.19})$$

where B is the integration constant.

The following considerations may be of use to estimate the constants γ and B . When the condition $f_F(E) \ll 1$ is satisfied, the fermion gas becomes nondegenerate. For such a gas, we can neglect unity in the left-hand side of Eq. (A.19) and rewrite the expression in the form:

$$f_F(E) = B^{-1} e^{-\gamma E}. \quad (\text{A.20})$$

Comparing Eq. (A.20) with Eq. (A.2), and keeping in mind Eq. (A.12), we obtain:

$$B = A^{-1} = e^{-\mu/(k_B T)}, \quad \gamma = -\beta = \frac{1}{k_B T}. \quad (\text{A.21})$$

Substituting into Eq. (A.19), we finally obtain

$$f_F(E) = \frac{1}{e^{(E-\mu)/(k_B T)} + 1}. \quad (\text{A.22})$$

A.3. Derivation of the Bose-Einstein distribution function

In contrast to fermions, bosons can occupy both vacant states and states already occupied by other bosons and they do it the more readily the greater is the occupancy of the states. Therefore, the rate of direct collisions $E_1 \rightarrow E_3, E_2 \rightarrow E_4$ will be the greater, the greater the numbers of particles in the initial states $f_{\text{Bose}}(E_1)$ and $f_{\text{Bose}}(E_2)$ and the higher the occupancy of the final states $f_{\text{Bose}}(E_3)$ and $f_{\text{Bose}}(E_4)$:

$$Q_d = c f_{\text{Bose}}(E_1) f_{\text{Bose}}(E_2) [1 + f_{\text{Bose}}(E_3)] [1 + f_{\text{Bose}}(E_4)]. \quad (\text{A.23})$$

The units in the brackets take account of the bosons' ability to go over not only to occupied states but to vacant states, for which $f_{\text{Bose}}(E_3) = f_{\text{Bose}}(E_4) = 0$, as well. For $f_{\text{Bose}}(E) \gg 1$ (the condition of nondegeneracy), the expression in the brackets in Eq. (A.23) becomes unity and Q_d becomes equal to the rate of direct collisions for the particles of a nondegenerate gas. For the rate of reverse collisions $E_3 \rightarrow E_1, E_4 \rightarrow E_2$, we obtain

$$Q_r = c f_{\text{Bose}}(E_3) f_{\text{Bose}}(E_4) [1 + f_{\text{Bose}}(E_1)] [1 + f_{\text{Bose}}(E_2)]. \quad (\text{A.24})$$

In the state of equilibrium $Q_d = Q_r$:

$$\begin{aligned} f_{\text{Bose}}(E_1) f_{\text{Bose}}(E_2) [1 + f_{\text{Bose}}(E_3)] [1 + f_{\text{Bose}}(E_4)] \\ = f_{\text{Bose}}(E_3) f_{\text{Bose}}(E_4) [1 + f_{\text{Bose}}(E_1)] [1 + f_{\text{Bose}}(E_2)]. \end{aligned} \quad (\text{A.25})$$

Dividing this expression by $f_{\text{Bose}}(E_1) f_{\text{Bose}}(E_2) f_{\text{Bose}}(E_3) f_{\text{Bose}}(E_4)$, we obtain

$$\left[\frac{1}{f_{\text{Bose}}(E_1)} + 1 \right] \left[\frac{1}{f_{\text{Bose}}(E_2)} + 1 \right] = \left[\frac{1}{f_{\text{Bose}}(E_3)} + 1 \right] \left[\frac{1}{f_{\text{Bose}}(E_4)} + 1 \right]. \quad (\text{A.26})$$

A comparison between Eqs. (A.26) and (A.3) shows that the function $1/f_{\text{Bose}}(E) + 1$ for bosons, satisfies the same equation as the function $f(E)$ for a nondegenerate gas. Therefore, we may make use of the result (A.10) writing it for bosons in the form

$$\frac{1}{[f_{\text{Bose}}(E)^{-1} + 1]} \frac{d}{dE} \left[\frac{1}{f_{\text{Bose}}(E)} + 1 \right] = \gamma. \quad (\text{A.27})$$

Integrating Eq. (A.27), we obtain

$$\frac{1}{f_{\text{Bose}}(E)} + 1 = B e^{\gamma E}. \quad (\text{A.28})$$

The values of the constants γ and B that enter this expression are the same as in the case of the degenerate fermion gas:

$$\gamma = \frac{1}{k_B T}, \quad B = e^{-\mu/(k_B T)}. \quad (\text{A.29})$$

Therefore,

$$f_{\text{Bose}}(E) = \frac{1}{e^{(E-\mu)/(k_B T)} - 1}. \quad (\text{A.30})$$

A.4. Tables

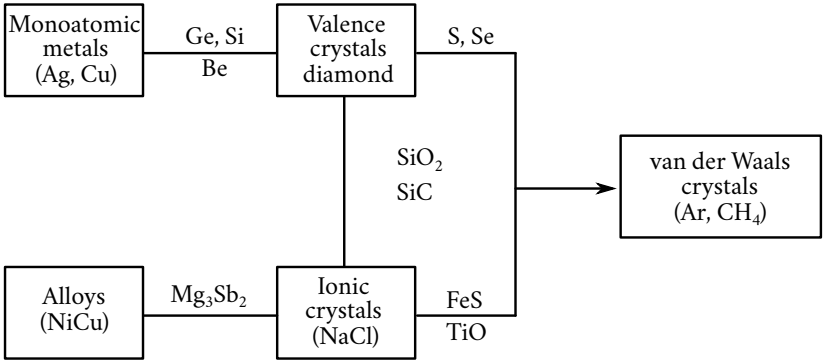
Table A.1

	I A	II A	III A	IV A	V A	VI A	VII A	VIII A	VIII B	VIII C	I B	II B	III B	IV B	V B	VI B	VII B	0	
I			Transition metals														H	He	
II	Li	Be												B	C	N	O	F	Ne
III	Na	Mg												Al	Si	P	S	Cl	Ar
IV	K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr	
V	Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe	
VI	Cs	Ba	La ¹	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn	
VII	Fr	Ra	Ac ²																
	I class											II class		III class			IV class		

¹ Lanthanides

² Actinides

Table A.2



Glossary of Symbols and Notations

A	lattice basis; Madelung constant; energy of exchange interaction
a	lattice constant; Bohr radius
B	magnetic field induction
C_V	constant volume heat capacity
C_e	heat capacity of electron gas
C	capacitance; Curie constant; Coulomb interaction energy
c	velocity of light
D	diffusion coefficient
D_n	electron diffusion coefficient
D_p	hole diffusion coefficient
d	barrier width; space charge layer width
E	energy; Young's modulus
E_a	acceptor excitation energy
E_b	bond energy
E_c	bottom of conduction band
E_d	donor excitation energy
E_e	electron gas energy
E_{exc}	excitation energy
$E_{\text{e.g}}$	energy gap
E_F	Fermi energy
E_g	forbidden energy band width
E_{lattice}	lattice energy
$E_{\text{n.m}}$	energy of normal mode
E_v	top of valence band
\mathcal{E}	electric field intensity
\mathcal{E}_c	contact field intensity

F	free energy; force
f	interaction force
$f(E)$	distribution function
$f_{\text{Bose}}(E)$	Bose-Einstein distribution function
$f_{\text{F}}(E)$	Fermi-Dirac distribution function
$f_{\text{M}}(E)$	Maxwell-Boltzmann distribution function
G	number of states; shear modulus
g	generation rate; anharmonicity coefficient; Lande factor
$g(E)$	density of states
$g(\omega)$	frequency distribution function of normal modes
H	magnetic field intensity
I	light flux; current
I_{b}	base current
I_{c}	collector current
I_{e}	emitter current
i	current density
i_{f}	forward current density
i_{e}	reverse current density
i_{s}	saturation current density
J	light intensity; exchange integral; intrinsic quantum number
J_{m}	magnetization
J_{s}	saturation magnetization
\mathbf{k}	wave vector of electron
k	absorption coefficient
k_{B}	Boltzmann constant
\mathcal{K}	heat conductivity
$\mathcal{K}_{\text{lattice}}$	lattice heat conductivity
\mathcal{K}_{e}	electron gas heat conductivity
L	linear dimension; diffusion length; Lorentz number
L_{n}	diffusion length of electrons
L_{p}	diffusion length of holes
l	orbital quantum number; transport mean free path
M	mass; electric dipole moment, magnetic moment
M_{J}	total magnetic moment of atom
m	electron rest mass; particle mass
m_{eff}	effective mass
m_{J}	magnetic quantum number of atom
m_{l}	magnetic quantum number of electron

m_n	electron effective mass
m_p	hole effective mass
N	number of particles
N_A	Avogadro's number
$N(E)$	total distribution function
N_a	acceptor concentration
N_d	donor concentration
N_{im}	concentration of impurities
n	concentration of particles; electron concentration in conduction band
n_i	equilibrium electron concentration in intrinsic semiconductor
n_{n0}	equilibrium majority carrier concentration in n-type semiconductor
n_{ph}	phonon concentration
n_{p0}	equilibrium minority carrier concentration in p-type semiconductor
P_j	atomic angular momentum
p	momentum; pressure; hole concentration
p_e	electron momentum
p_F	Fermi energy electron momentum
p_i	hole concentration in intrinsic semiconductor
p_l	orbital angular momentum
p_{n0}	equilibrium minority carrier concentration in n-type semiconductor
p_{ph}	phonon momentum
p_{p0}	equilibrium majority carrier concentration in p-type semiconductor
p_s	electron spin
Q	quantity of heat; electric charge
Q_d	destruction energy
Q_p	Peltier heat
Q_s	sublimation energy
q	electron charge; phonon wave number
R	recombination rate; gas constant
R_H	Hall coefficient
R_n	electron recombination rate
R_p	hole recombination rate
r	distance between particles
r_0	equilibrium interparticle distance
S	entropy; slip plane
T	temperature
T_C	Curie temperature
T_{cr}	critical or transition temperature

T_F	Fermi temperature
T_i	intrinsic conductivity transition temperature
T_s	impurity exhaustion temperature
U	potential energy
U_b	bond energy
U_m	magnetic energy
u	mobility
u_n	electron mobility
u_p	hole mobility
V	potential; voltage; tension; volume
V_c	contact potential difference; collector voltage
V_{ph}	photo-emf
V_t	thermal emf
v	velocity
v_o	thermal velocity
v_d	drift velocity
v_e	electron velocity
v_F	Fermi energy electron velocity
W	work
x	particle's displacement from equilibrium position
α	polarizability of molecules; free surface energy; linear thermal expansion coefficient; temperature coefficient of resistance; differential, or specific, thermoelectric power
β	bond rigidity coefficient; quantum efficiency (yield)
Γ	phase volume
γ	relative shear deformation, gyromagnetic ratio
ε	relative permittivity; quantum energy; relative extension deformation
ε_{ph}	phonon energy
Θ	Debye temperature; paramagnetic Curie point
Θ_C	ferromagnetic Curie point
λ	wavelength; mean free path; magnetic field penetration depth in conductor
λ_{ph}	phonon mean free path
μ	chemical potential (Fermi level); molecular mass; magnetic susceptibility; magnetic moment
μ_B	Bohr magneton
μ_i	electron orbital magnetic moment
μ_n	Fermi level in n-type semiconductor

μ_p	Fermi level in p-type semiconductor
μ_s	electron intrinsic magnetic moment
ν	frequency; number of collisions; Poisson's coefficient
Π	Peltier coefficient
σ	specific resistance; space charge density; specific conductance; normal stress; theoretical strength
σ_{dif}	differential conductivity
σ_i	intrinsic specific conductance
σ_{im}	impurity (extrinsic) specific conductance
σ_{ph}	photoconductivity, phonon effective cross section
σ_{ph0}	stationary photoconductivity
σ_r	real (technical) strength
τ	lifetime; relaxation time; tangential (shear) stress; durability
τ_{cr}	critical shear stress
τ_n	electron "lifetime"
τ_p	hole lifetime
φ	electron potential energy; angle
φ_0	equilibrium potential barrier in p-n junction
χ	thermodynamic work function; magnetic susceptibility
ψ	wave function
ω	angular frequency
ω_D	Debye frequency
ω_L	Larmor frequency

Bibliography

General

- 0.1.** V. L. Bonch-Bruevich and S. G. Kalashnikov: Semiconductor Physics, "Nauka", Moscow (1977) (in Russian).
- 0.2.** G. I. Epifanov: Physical Principles of Microelectronics, Mir Publishers, Moscow (1974).
- 0.3.** P. T. Oreshkin: Physics of Semiconductors and Dielectrics, "Vysshaya Shkola", Moscow (1977) (in Russian).
- 0.4.** L. S. Stil'bans: Semiconductor Physics, "Sovetskoe radio", Moscow (1967) (in Russian).
- 0.5.** K. V. Shalimova: Semiconductor Physics, "Energiya", Moscow (1976) (in Russian).
- 0.6.** V. V. Novikov: Theoretical Foundation of Microelectronics, "Vysshaya shkola", Moscow (1972) (in Russian).
- 0.7.** J. M. Ziman: Principles of the Theory of Solids (2nd edition), Cambridge U.P., London (1970).
- 0.8.** R. L. Sproull: Modern Physics (2nd edition), Wiley, New York (1964).
- 0.9.** K. Seeger: Semiconductor Physics, Springer, Wien (1973).

Chapter 1

- 1.1.** G. C. Pimentel and R. D. Spratley: Chemical Bonding Clarified Through Quantum Mechanics, Holden-Day, San Francisco (1969).
- 1.2.** W. Haberditzl: Bausteine der Materie und chemische Binding, Deutscher Verlag der Wissenschaften, Berlin (1972).
- 1.3.** L. Pauling: General Chemistry (3rd edition), W. H. Freeman, San Francisco (1970).
- 1.4.** J. A. Campbell: Chemical Systems, W. H. Freeman, San Francisco (1970).
- 1.5.** J. Ficini, N. Lumbroso-Bader, and J.-C. Depeyay: Elements de chimiephysique, Hermann, Paris (1968-69).
- 1.6.** C. A. Wert and R. M. Thomson: Physics of Solids (2nd edition), McGraw-Hill, New York (1970).

Chapter 2

- 2.1.** Ya. I. Frenkel: Introduction to the Theory of Metals (4th edition), "Nauka", Moscow (1972) (in Russian).
- 2.2.** M. Kh. Rabinovich: Strength and Superstrength of Metals, Izd-vo Akad. Nauk SSSR, Moscow (1963) (in Russian); Strength, Temperature, Time, "Nauka", Moscow (1968) (in Russian).

- 2.3.** A. V. Stepanov: Fundamentals of the Theory of Practical Strength of Crystals, "Nauka", Moscow (1974) (in Russian).
- 2.4.** G. M. Bartenev and Yu. S. Zuev: Strength and Rupture of Hyperelastic Materials, "Khimiya", Moscow-Leningrad (1964) (in Russian).
- 2.5.** Physical Metallurgy, R. W. Cahn (Ed.), North-Holland, Amsterdam (1965).
- 2.6.** Fracture, H. Liebowits (Ed.), Academic Press, New York (1972).
- 2.7.** F. A. McCertrrock and A. S. Argon: Mechanical Behaviour of Materials, Addison-Wesley, Reading, Mass. (1966).
- 2.8.** See 1.6.

Chapter 3

- 3.1.** V. G. Levich: Introduction to Statistical Physics (2nd edition), GITTL, Moscow (1954) (in Russian).
- 3.2.** F. Reif: Statistical Physics (Berkeley Physics Course, Vol. 5), McGraw-Hill, New York (1972).
- 3.3.** H. Bethe and A. Sommerfeld: Electronentheorie der Metalle, Springer, Berlin (1933).
- 3.4.** C. Kittel: Introduction to Solid State Physics (4th edition), Wiley, New York (1971); Elementary Solid State Physics: A Short Course, Wiley, New York (1962).
- 3.5.** See 0.1-0.7.

Chapter 4

- 4.1.** A. F. Ioffe: Physics of Semiconductors, Infosearch, London (1960).
- 4.2.** A. J. Dekker: Electrical Engineering Materials, Prentice-Hall, Englewood Cliffs, N. J. (1959).
- 4.3.** G. A. Slack: "Heat conduction in solids" in Encyclopaedic Dictionary of Physics, J. Thewlis (Ed.) Vol. 3, Pergamon Press, Oxford (1961), 601-610.
- 4.4.** See 0.1-0.8.

Chapter 5

- 5.1.** J. Callaway: Energy Band Theory, Academic Press, New York (1964).
- 5.2.** Semiconductors, N. B. Hannay (Ed.), Reinhold, New York (1959).
- 5.3.** R. A. Smith: Semiconductors, Cambridge U.P., London (1959).
- 5.4.** J. N. Shive: The Properties, Physics, and Design of Semiconductor Devices, Van Nostrand, Princeton, N.J. (1959).
- 5.5.** See 0.1-0.9, 1.6, 3.3, 4.1.

Chapter 6

- 6.1.** G. Mirdel: Elektrophysik, Berlin (1970). **6.2.** A. C. Rose-Innes and E. H. Rhoderick: Introduction to Superconductivity, Pergamon Press, Oxford (1969).
- 6.3.** J. Bardeen and J. R. Schrieffer: "Recent developments in superconductivity", in Progr. Low Temp. Phys., Vol. 3, Chapter 4, North-Holland, Amsterdam (1961).
- 6.4.** D. H. Douglass and L. M. Falikov: "The superconductivity energy gap", in Progr. Low Temp. Phys., Vol. 4, Wiley, New York (1964), 97.
- 6.5.** D. N. Langenberg, D. J. Scalapino, and B. N. Taylor: "The Josephson effect", Sci. Amer. 214, No. 5 (1966), 30-9.

- 6.6. J. Berdeen: *Physics Today* 26, No. 7 (1973).
- 6.7. J. Schrieffer: *Physics Today* 26, No. 7 (1973).
- 6.8. L. Cooper: *Physics Today* 26, No. 7 (1973).
- 6.9. J. Giaver: *Science* 183, No. 4131 (1974).
- 6.10. B. D. Josephson: *Science* 184, No. 4136 (1974).
- 6.11. W. P. Jolly: *Cryoelectronics*, The English U.P., London (1972).
- 6.12. See 0.1-0.9, 1.6, 2.1, 3.3, 4.1, 4.2, 5.2-5.4.

Chapter 7

- 7.1. S. V. Vonsovskii: *The Contemporary Science of Magnetism*, Gostekhizdat, Moscow-Leningrad (1953) (in Russian).
- 7.2. Ya. G. Dorfman: *Magnetic Properties and Structure of Matter*, Gostekhizdat, Moscow (1955) (in Russian).
- 7.3. L. V. Kirenskii: *Magnetism*, Izd-vo Akad. Nauk SSSR, Moscow (1963) (in Russian).
- 7.4. See 0.1-0.8, 1.6, 2.1, 3.3, 4.2.

Chapter 8

- 8.1. Ya. I. Fedotov: *Fundamentals of Physics of Semiconductor Devices*, "Sovetskoe radio", Moscow (1969) (in Russian).
- 8.2. I. P. Zharebtsov: *Electronics* (reprint of the 2nd edition), Mir Publishers, Moscow (1975).
- 8.3. V. V. Pasynkov, L. K. Chirkin, and A. D. Shinkov: *Semiconductor Devices*, "Vysshaya shkola", Moscow (1966) (in Russian).
- 8.4. A. F. Gorodetskii and A. F. Kravchenko: *Semiconductor Devices*, "Vysshaya shkola", Moscow (1967) (in Russian).
- 8.5. S. N. Levine: *Principles of Solid-State Microelectronics*, Holt, Rinehart and Winston, New York (1963).
- 8.6. J. H. Kalish: *Microminiature Electronics*, Bobbs-Merrill, Indianapolis (1967).
- 8.7. G. F. Alfrey: *Physical Electronics*, Van Nostrand, Princeton, N.J. (1964).
- 8.8. See 0.1-0.9, 1.6, 4.1, 5.2-5.4, 6.1.

Chapter 9

- 9.1. See 0.1-0.7, 0.9, 3.3, 4.1, 5.3, 5.4, 6.1.

OTHER MIR TITLES

Fundamentals of the Theory of Electricity

I. TAMM, Mem. USSR Acad. Sci.

The present book is mainly intended for students of physical faculties of universities who have mastered differential and integral calculus and vector algebra; the fundamentals of vector analysis are set out in the text as needed. The main object of this course is the determination of the physical meaning and content of the main laws of the theory of electricity. Although not treating the technical applications of the theory, the author prepares the reader as far as possible to a direct transition to studying the applied theory of electricity. The book contains a number of problems forming an organic part of the text. The solutions of the majority of them are needed for an understanding of the subject matter.

Special Theory of Relativity

V. UGAROV, Cand. Sc.

This is the English translation of the second revised and enlarged edition of the book which was first published in 1969 under the same title. It is written as a textbook to be used both by university students and teachers as well as high school teachers. Although the general layout of the book has not been changed, the principles of the theory are given in more details now and the greater attention is paid to the four-dimensional treatment. Different ways of presenting the special theory of relativity are described. Sections devoted to the methodology and history of the special theory of relativity are added. The chapter on electrodynamics is enlarged. Finally the paper written by Academician V. L. Ginzburg and entitled "Who Created the Special Theory of Relativity and How It Was Done" is included.

To the Reader

Mir Publishers welcome your comments on the content, translation and design of this book.

We would also be pleased to receive any proposals you care to make about our future publications.

Our address is:

USSR, 129820, Moscow 1-110, GSP,
Pervy Rizhsky Pereulok, 2
Mir Publishers

Printed in the Union of Soviet Socialist Republics

